

Remarks

Claims 57-82 are pending in the subject application. By this Amendment, Applicants have amended claims 62 and 73 to correct inadvertent errors. Entry and consideration of the amendment presented herein is respectfully requested. Accordingly, claims 57-82 are currently before the Examiner. Favorable consideration of the pending claims is respectfully requested.

As an initial matter, replacement Figures 3 and 23 are submitted herewith. Replacement figures were necessary in order to remove any indication of color. No new matter has been added by these amendments. Entry of the replacement drawings is respectfully requested.

The application is objected to on the grounds that the subject specification fails to comply with the requirements of 37 CFR 1.821(a)(1) and (a)(2). Specifically, the Examiner indicates that a new sequence listing is required because the sequences in Figures 1-6, 13-15 and 22-24 are not identified by a sequence identifier number. The sequences shown in Tables III, IV and V of the subject specification have also been designated with a sequence identifier number. A Submission of Sequence Listing Under §1.821, including a replacement sequence listing on paper and a computer readable format, is attached. Accordingly, reconsideration and withdrawal of the objection is respectfully requested.

The disclosure is objected to because it contained embedded hyperlinks or other forms of browser executable code. Applicants respectfully submit that this issue is moot in view of the amendments made to the specification. Accordingly, reconsideration and withdrawal of the objection is respectfully requested.

Claims 57-82 are rejected under 35 U.S.C. § 101 on the grounds that the claimed invention lacks a substantial utility. In addition, claims 57-82 are rejected under 35 U.S.C. §112, first paragraph, as nonenabled on the grounds that the subject specification fails to teach a substantial utility for the claimed invention and, therefore, an ordinarily skilled artisan would not know how to use the claimed invention. The Office Action argues that the specification states that the invention is based upon the identification of an Open Reading Frame (ORF) in the human genome encoding a novel Preadipocyte factor-1-like polypeptide (referred to as SCS0009 and other splice variants. The Office Action further argues that the asserted utilities of the claimed polypeptides have been based

upon domain organization and that the asserted functions of the polypeptides are merely hypothetical and based upon domain homology. The Office Action further argues that function cannot be predicted based upon structural similarity to a protein in the sequence databases, citing to Skolnick *et al.*, *Trends Biotechnol.* (2000) 18:34-39. Applicants respectfully assert that the claimed invention has substantial utility and, therefore, is enabled and traverse this rejection.

The examiner bears the initial burden of showing that a claimed invention lacks patentable utility. *See In re Brana*, 51 F.3d 1560, 1566, 34 USPQ2d 1436, 1441 (Fed. Cir. 1995) (“Only after the PTO provides evidence showing that one of ordinary skill in the art would reasonably doubt the asserted utility does the burden shift to the applicant to provide rebuttal evidence sufficient to convince such a person of the invention’s asserted utility.”). The Patent Office must also articulate the factual assumptions and provide evidentiary support relied upon in establishing the *prima facie* showing. Applicants also note that compliance with the utility requirement of 35 U.S.C. § 101 and the enablement requirement of 35 U.S.C. § 112, first paragraph, does not turn on whether an example is disclosed (see M.P.E.P. §2164.02). Indeed, the lack of working examples or methods that indicate that the claimed polypeptide is involved in any of the asserted activities, standing alone, cannot be the basis for a rejection under 35 U.S.C. 101 or 35 U.S.C. 112. *Tol-O-Matic, Inc. v. Proma Produkt-Und Mktg. Gesellschaft m.b.h.*, 945 F.2d 1546, 1553, 20 USPQ2d 1332, 1338 (Fed. Cir. 1991).

Applicants also note that the as-filed specification indicates (at page 5, lines 16-26):

The novel polypeptide described herein was identified on the basis of a consensus sequence for human Preadipocyte factor-1-like polypeptides in which the number and the positioning of selected amino acids are defined for a protein sequence having a length comparable to known Preadipocyte factor-1-like polypeptides.

The totality of amino acid sequences obtained by translating the known ORFs in the human genome were challenged using this consensus sequence, and the positive hits were further screened for the presence of predicted specific structural and functional “signatures” that are distinctive of a polypeptide of this nature, and finally selected by comparing sequence features with known Preadipocyte factor-1-like polypeptides. Therefore, the novel polypeptides of the invention can be predicted to have Preadipocyte factor-1-like activities.

Furthermore, the parameters used to identify the sequences and pertinent domains are provided in

Examples 1 and 4 and the Office Action fails to establish any reason one skilled in the art would doubt the asserted utilities of the claimed polypeptides or the functions assigned to the claimed polypeptides on the basis of the information disclosed in the as-filed application.

Applicants now turn to Patent Office's arguments that the teaching of Skolnick *et al.* support a finding that the claimed invention lacks patentable utility. With regard to the teachings of that reference, Applicants submit that its teachings are not as absolute as argued in the Office Action. Applicants submit, herewith, a Journal of Molecular Biology article (Wilson *et al.*, 2000, *J. Mol. Biol.*, Vol. 297, pp. 233-249) published in the same timeframe as Skolnick *et al.* As opposed to the broad and sweeping generalizations found in Skolnick *et al.*, Wilson *et al.* specifically analyzed and addressed the degree to which structural annotation can be transferred between sequences at a given level of sequence similarity (see Practical Implications and Figure 7, pages 245-247). As noted by Wilson *et al.*, if a sequence matches a protein database structure with an *e*-value of 0.001 and a percent identity of 30%, the polypeptide is virtually certain to have the same fold, the polypeptide has 66% likelihood of having the same exact function, and that the proteins have about a 99% chance of having the same functional class. In this case, the claimed polypeptide has 34% identity, 46% similarity and an *e*-value of $2e-47$ (see attached alignment for SEQ ID NO: 2 and PREF-1 (SEQ ID NO: 38)). Thus, Wilson *et al.* would indicate that it is more likely than not that the two polypeptides would have the same or similar function and it is respectfully submitted that those skilled in the art would not have had a reason to doubt the asserted utility of the claimed invention. Thus, it is respectfully submitted that the claimed invention has a specific, credible and substantial utility and that one skilled in the art would know how to use the claimed invention in view of the teachings of the specification. Accordingly, reconsideration and withdrawal of the rejections is respectfully requested.

Claims 57, 64-69 and 77-82 are rejected under 35 U.S.C. § 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. Applicants respectfully assert that there is adequate written description in the subject specification to convey to the ordinarily skilled artisan that they had possession of the

claimed invention. The Office Action argues that the as-filed specification fails to provide adequate written description and evidence of possession for the claimed genus of variants. The Office Action also notes that the only factor present in the claims is the recitation of an activity (“prevents the terminal differentiation of preadipocytes”) and that there is no identification of any portion of the structure that must be conserved for the required activity. The Office Action further argues that it is unclear as to what molecules are within the genus of “active variants” as the specification does not provide a complete or partial structure and fails to provide a representative number of species for the recited genus. Applicants traverse the rejection.

At the outset, Applicants note that the claims are directed to an isolated polypeptide comprising an active variant of SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 8, SEQ ID NO: 9 or SEQ ID NO: 10, wherein any amino acid specified in the sequence is non-conservatively substituted, provided that no more than 15% of the amino acid residues are substituted and said active variant prevents the terminal differentiation of preadipocytes. Thus, and contrary to the assertion in the Office Action, the claims provide at least a partial structure of the claimed polypeptide variants; namely, any one of SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 8, SEQ ID NO: 9 or SEQ ID NO: 10 in which no more than 15% of the amino acid residues are substituted and wherein the polypeptide retains the ability to prevent the terminal differentiation of preadipocytes.

Turning to the argument that the as-filed specification fails to identify the portion of the molecule that is associated with the ability to prevent the terminal differentiation of preadipocytes, Applicants respectfully submit that the domain of PREF-1 associated with that activity was known to those skilled in the art (see, for example, Smas *et al.* (*Mol. Cell. Biol.*, 1997, Vol. 17, pp. 977-988, a copy of which is attached). As discussed in that publication, the ability to prevent terminal differentiation of preadipocytes is associated with the N-terminal region of the polypeptide and soluble PREF-1 containing that domain inhibits preadipocytes differentiation (see pages 981-983). As the Patent Office is aware, the Federal Circuit has made clear that the specification need not describe every permutation of an invention nor subject matter known to those of skill in the art. *Capon v. Eshhar*, 418 F.3d 1349,1359-60 (Fed. Cir. 2005). Moreover, an adequate written

description of an invention that involves biological macromolecules need not contain a recitation of each known structure, particularly when those structures are already known in the art. *Falkner v. Inglis*, 448 F.3d 1357, 1366 (Fed. Cir. 2006) (“the forced recitation of known sequences in patent disclosures would only add unnecessary bulk to the specification. Accordingly, we hold that where . . . accessible literature sources clearly provided, as of the relevant date, [the sequences], satisfaction of the written description requirement does not require either the recitation or incorporation by reference”). Accordingly, it is respectfully submitted that the as-filed specification and currently pending claims fully comply with the written description requirement and reconsideration and withdrawal of the rejection under 35 U.S.C. § 112, first paragraph, is respectfully requested.

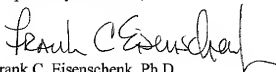
It should be understood that the amendments presented herein have been made solely to expedite prosecution of the subject application to completion and should not be construed as an indication of Applicants’ agreement with or acquiescence in the Examiner’s position. Applicants expressly reserve the right to pursue the invention(s) disclosed in the subject application, including any subject matter canceled or not pursued during prosecution of the subject application, in a related application.

In view of the foregoing remarks and amendments to the claims, Applicants believe that the currently pending claims are in condition for allowance, and such action is respectfully requested.

The Commissioner is hereby authorized to charge any fees under 37 CFR §§1.16 or 1.17 as required by this paper to Deposit Account No. 19-0065.

Applicants invite the Examiner to call the undersigned if clarification is needed on any of this response, or if the Examiner believes a telephonic interview would expedite the prosecution of the subject application to completion.

Respectfully submitted,



Frank C. Eisenschenk, Ph.D.

Patent Attorney

Registration No. 45,332

Phone No.: 352-375-8100

Fax No.: 352-372-5800

Address: P.O. Box 142950

Gainesville, FL 32614-2950

FCE/jb/sl

Attachments: Annotated and Replacement Figures 3 and 23
Submission of Sequence Listing and Statement
New pages 1-37 (Sequence Listing)
Wilson *et al.*, 2000, *J. Mol. Biol.*, Vol. 297, pp. 233-249
Alignment for SEQ ID NO: 2 and PREF-1 (SEQ ID NO: 38)
Smas *et al.*, *Mol. Cell. Biol.*, 1997, Vol. pp. 977-988

Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure and Function through Traditional and Probabilistic Scores

Cyrus A. Wilson¹, Julia Kreychman¹ and Mark Gerstein^{1,2*}

¹Department of Molecular Biophysics and Biochemistry

²Department of Computer Science, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, CT 06520, USA

Measuring in a quantitative, statistical sense the degree to which structural and functional information can be “transferred” between pairs of related protein sequences at various levels of similarity is an essential prerequisite for robust genome annotation. To this end, we performed pairwise sequence, structure and function comparisons on ~30,000 pairs of protein domains with known structure and function. Our domain pairs, which are constructed according to the SCOP fold classification, range in similarity from just sharing a fold, to being nearly identical. Our results show that traditional scores for sequence and structure similarity have the same basic exponential relationship as observed previously, with structural divergence, measured in RMS, being exponentially related to sequence divergence, measured in percent identity. However, as the scale of our survey is much larger than any previous investigations, our results have greater statistical weight and precision. We have been able to express the relationship of sequence and structure similarity using more “modern scores,” such as Smith-Waterman alignment scores and probabilistic *P*-values for both sequence and structure comparison. These modern scores address some of the problems with traditional scores, such as determining a conserved core and correcting for length dependency; they enable us to phrase the sequence-structure relationship in more precise and accurate terms. We found that the basic exponential sequence-structure relationship is very general: the same essential relationship is found in the different secondary-structure classes and is evident in all the scoring schemes. To relate function to sequence and structure we assigned various levels of functional similarity to the domain pairs, based on a simple functional classification scheme. This scheme was constructed by combining and augmenting annotations in the enzyme and fly functional classifications and comparing subsets of these to the *Escherichia coli* and yeast classifications. We found sigmoidal relationships between similarity in function and sequence, with clear thresholds for different levels of functional conservation. For pairs of domains that share the same fold, precise function appears to be conserved down to ~40% sequence identity, whereas broad functional class is conserved to ~25%. Interestingly, percent identity is more effective at quantifying functional conservation than the more modern scores (e.g. *P*-values). Results of all the pairwise comparisons and our combined functional classification scheme for protein structures can be accessed from a web database at <http://bioinfo.mbb.yale.edu/align>

© 2000 Academic Press

Keywords: bioinformatics; sequence similarity; percent identity; structure similarity; functional classification

*Corresponding author

Abbreviations used: EC, Enzyme Commission; EST, expressed sequence tags; SCOP, structural classification of proteins; GO, Gene Ontology Project.

E-mail address of the corresponding author: Mark.Gerstein@yale.edu

Introduction

The problem of genome annotation

Perhaps the most valuable information to be gained from a genome analysis is functional annotation of all the gene products. Unfortunately, of all the proteins whose sequences are known, functions have been experimentally determined for only a very small number (Andrade & Sander, 1997). Given the current size and accessibility of sequence and structure data, homologs of a newly sequenced gene's product can be identified *via* database searches, and probable structure and function assigned to the gene product (Bork *et al.*, 1998). This is based on the concept that sequence similarity implies structural and functional similarity. However, structural and functional annotations should be transferred with caution. If a protein is assigned an incorrect function in a database, the error could carry over to other proteins for which structure or function is inferred by homology to the errant protein (Brenner, 1999; Karp, 1996, 1998a). In large databases such an error can propagate out of control, presenting a serious quality control issue as we move to larger genomes from multicellular organisms.

Benchmarking fold and function recognition

Here, we used manually curated structural and functional classifications as standards in analyzing to what degree annotations of a protein's structure and function can be transferred to a similar sequence. The knowledge gained from the study can be used to establish confidence levels for structure and function prediction, improving our understanding of how long it will take to annotate accurately an entire genome.

Our simultaneous analysis of relationships between sequence and structure, sequence and function, and structure and function (Figure 1) may provide insight into paradigms for functional prediction other than that based alone on sequence similarity (Enright *et al.*, 1999).

Past results

Sequence-structure

The transfer of structural annotation is well characterized. Chothia & Lesk (1986, 1987) found that structural divergence, when expressed in terms of the RMS separation of matching alpha carbon atoms, was an exponential function of sequence divergence, expressed in terms of the fraction of residues that differed between sequences. The reliability of structural annotation transferred by homology, then, depends on the sequence identity of the homologous proteins (Chothia & Lesk, 1986). Flores *et al.* (1993), Russell & Barton (1994), and Russell *et al.* (1997) observed the same general trend, and also characterized the conservation of structural features other than the

C α backbone, such as secondary structure, accessibility and torsion angles. A paper by Wood & Pearson (1999) re-expressed the sequence-structure relationship in terms of statistically based "Z-scores" and found that this relationship had a simple linear form in terms of these scores. They also noted that protein families differed in detail in the slope of this linear relationship.

Others have focused on the limits of sequence comparison, specifically around the "twilight zone," the region of sequence similarity that does not reliably imply structural homology (Doolittle, 1987), and on establishing cut-offs for significant sequence similarity. Using the SCOP structural classification (Murzin *et al.*, 1995), Brenner *et al.* (1998) benchmarked the effectiveness of the popular FASTA and BLASTP programs and their probabilistic scoring schemes (i.e. the *e*-value) (Pearson & Lipman, 1988; Pearson, 1996; Altschul *et al.*, 1990, 1994; Karlin & Altschul, 1993). They found that in making fold assignments, the FASTA *e*-value closely tracked the number of false positives, i.e. the error rate, and that at a conservative *e*-value cut-off of 0.001, the FASTA program could detect nearly all the relationships that would be detected by a full Smith-Waterman comparison (Smith & Waterman, 1981). Specifically, they found that FASTA with a 0.001 threshold would find 16% more of the structural relationships in SCOP than would be found by standard sequence comparison with a 40% identity threshold. This rigorous benchmarking approach has been extended to assess transitive sequence comparison, through a third intermediate sequence and multiple-sequence matching programs such as PSI-blast (Park *et al.*, 1997, 1998; Gerstein, 1998a; Salamov *et al.*, 1999). In a related study Rost (1999) worked on characterizing the region after the twilight zone, which he called the "midnight zone". In a sense these benchmarking studies have culminated in the CASP fold recognition experiments (Moult *et al.*, 1997; Sternberg *et al.*, 1999).

Sequence-function

Although the exact dependence of functional similarity on sequence and structural similarity is not completely clear, initial indications of a gene product's function are most often based on simple sequence similarity (Bork *et al.*, 1994, 1998). Often these are merely based on the best hit in database comparisons; see, for example, the annotation of some of the early genomes (Fraser *et al.*, 1995, 1998). However, possibilities for more robust annotation transfer are increasingly available. One looks at the pattern of hits amongst different phylogenetic groups (Tatusov *et al.*, 1997). Often these focus on the existence of key motifs and patterns associated with function (Zhang *et al.*, 1998; Bork & Koonin, 1996; Attwood *et al.*, 1999).

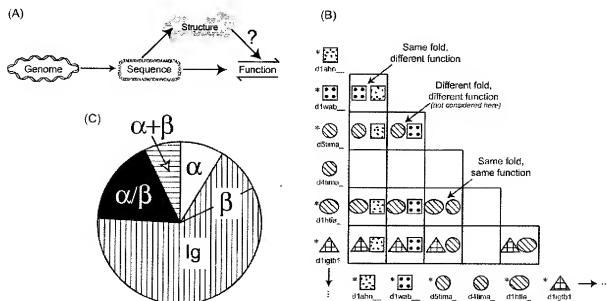


Figure 1. This Figure schematically depicts certain aspects of our comparison methodology. (a) The paradigm relating sequence to structure to function. There has not been as much assessment of functional annotation transfer based on structure as there has been with sequence-based structural and functional annotation transfer. (b) How we conceptualized our analysis in terms of pairs. A few examples of SCOP domains (identified on the left and bottom) are included from our comparison. In the Figure the shape represents fold, and the pattern represents function. We have highlighted some example categories of pairs: a pair that shares fold and function, a pair that shares fold but not function and a pair that shares neither fold nor function. The latter category of pairs is not considered in our investigation; we looked only at paired domains with the same fold. In constructing our pairs, we used only a representative set of SCOP domains. This is illustrated in the Figure by the domains flagged with asterisks. Note, in particular, that the SCOP domain d4tma₁ is not paired with anything because it is represented by d5tma₁, which is the same species and protein. For each level of pairs (fold, superfamily, family), cluster representatives were chosen for the level below: (i) for family pairs, one representative was selected from each species/protein, the level below, and then paired with all the other representatives within its family; (ii) for superfamily pairs, one representative was chosen from each family, unless there were domains in the family that shared less than 40% sequence identity, in which case additional representatives were included, each not more than 40% identical with the other representatives from the family (this occurs, for instance, for the globins); and (iii) likewise for fold pairs, one representative was chosen from each superfamily, more if there were domains with less than 40% sequence identity. (c) Subdivides the pairs into the four SCOP classes from which they were composed: (i) all- α , domains consisting of α -helices; (ii) all- β , domains consisting of β -sheets; (iii) α/β , domains with integrated α -helices and β -strands; and (iv) $\alpha + \beta$, domains with segregated α -helices and β -strands. We initially set apart the immunoglobulins from the rest of the all- β pairs because we realized that their large number biases our data. However, we compared the results for the immunoglobulin pairs to all other pairs and found that they generally exhibit the same behavior as the other pairs. Therefore we decided to leave them in the comparison.

Sequence-structure-function

One way that the better-defined sequence-structure relationship can assist in function prediction is initially to predict the structure of an uncharacterized sequence and then predict the function based on the limited repertoire of functions known to occur with that structure. To some degree this was achieved by Fetrow and co-workers (Fetrow *et al.*, 1998; Fetrow & Skolnick, 1998). They predicted structural profiles based on threading and *ab initio* methods, and then searched with these against profiles of known structures in order to predict function.

In related work, Russell *et al.* (1998) discussed using identification of structural binding sites in

predicting protein function. In a comprehensive study, Hegyi & Gerstein (1999) investigated to what degree folds were associated with functions. They found that most folds were associated with one or two functions with the exception of a few special folds, such as the TIM barrel, that could carry out numerous functions. Furthermore, they found that particular folds were often confined to distinct phylogenetic groups, an additional fact that can feed into an integrated sequence-structure-function analysis (Gerstein & Hegyi, 1998; Gerstein, 1997, 1998b,c).

Here, we look at pairwise comparisons of protein sequence, structure and function among proteins that share the same fold. We assess the

trends relating sequence, structure and function and consider the implications for structural and functional annotation transfer.

New developments: probabilistic scoring and growth of the databank

The past studies regarding sequence, structure and function relationships often used RMS separation and percent sequence identity (or a linear variant of it, such as the fraction of mutated residues) to express similarities in structure and in sequence, respectively. However, it has become increasingly common to use probabilistic scoring schemes (*P*-values) to express the quality of a match in terms of statistical significance rather than an arbitrary raw score such as percent identity (Pearson, 1998; Karlin & Altschul, 1990, 1993; Karlin *et al.* 1991; Altschul *et al.* 1994; Bryant & Altschul, 1995; Abagyan & Batalyov, 1997). With *P*-values, scores from different investigations can be compared in a common framework. Recently, it was found that sequence and structure similarity significance can be expressed as *P*-values in the same unified statistical framework (Levitt & Gerstein, 1998). Here, we use such probabilistic scoring methods to overcome the limitations of the more traditional scores.

Another recent development is the tremendous growth in the number of solved structures. The RCSB Protein Data Bank (Bernstein *et al.* 1977) now contains more than 10,000 protein structures. These structures are broken into more than 18,000 domains, and then domains that share a fold are paired up with each other for comparison (Figure 1(b)). Here, we survey ~30,000 pairs of protein domains that are known to have the same fold, approximately 1000 times the number compared by Chothia & Lesk (1986). The large scale of this comparison affords greater statistical weight to the results.

Alignment of 30,000 pairs from SCOP

The basic unit of comparison: a pair of protein domains

The protein domains that we studied were classified by SCOP, a Structural Classification of Proteins (Murzin *et al.* 1995; Brenner *et al.* 1996; Hubbard *et al.* 1997), a hierarchy of five levels: (i) class, domains that have the same secondary structural content (all- α , all- β , α/β , or $\alpha + \beta$); (ii) fold, domains that geometrically share the same tertiary fold; (iii) superfamily, domains descended from the same ancestor (but which lack measurable sequence similarity); (iv) family, domains in the same protein sequence family (which have appreciable sequence similarity); and (v) species and protein.

Pairs of protein domains that are grouped together at the fold, superfamily or family level form the basic unit of our comparisons.

Selection of pairs

There is potentially a huge number of pairs of domains that can be constructed out of the relationships in SCOP. For instance, in the current version of SCOP there are ~3.9 million potential pairs between domains sharing the same fold. Most of these are between nearly identical structures. In order to keep the number of pairs manageable, we used a straightforward clustering scheme, described in the legend to Figure 1. We selected 29,454 representative pairs from the total in SCOP. To achieve a wide range of similarities, we constructed the pairs on three levels of the SCOP hierarchy: (i) family pairs, 19,542 pairs of domains in the same family; (ii) superfamily pairs, 4220 pairs of domains in the same superfamily but different families; and (iii) fold pairs, 5692 pairs of domains in the same fold but different superfamilies.

All the selected domains were at least 50 residues in length and were drawn from the four major SCOP secondary-structural classes: all- α , all- β , α/β , and $\alpha + \beta$ (Figure 1(c)).

We automatically aligned each of our selected domain pairs twice, once by global Needleman-Wunsch sequence comparison (Needleman & Wunsch, 1971; Myers & Miller, 1998) and then by structure (Gerstein & Levitt, 1996, 1998), calculating scores for sequence and structural similarity.

Web-accessible database

The results of all the pairwise comparisons are available via a searchable database on the web at <http://bioinfo.mbb.yale.edu/align>. The query engine allows searches of individual SCOP pairs, all pairs that include a given SCOP domain, or all pairs containing any SCOP domain contained in a given PDB entry.

Traditional scores: RMS and percent identity

The sequence-structure relation, as expressed by the root-mean-square (RMS) of the aligned C α distances and percent sequence identity, has been previously characterized as an exponential function by Chothia & Lesk (1986) and others (Flores *et al.* 1993; Russell & Barton, 1994; Russell *et al.* 1997). As Figure 2 illustrates, our data display a similar trend. (Exact equations are given in the legend to Figure 2.) However, we have one thousand times as many data points as in Chothia and Lesk's original study (30,000 as opposed to 30).

The main difference between our results and the previous studies is due to differences in RMS "trimming" methods. By trimming we refer to the process of removing the worst-fitting aligned atoms from the RMS calculation, to arrive at a structural "core." This was first developed in Lesk's sieve-fit procedure (Lesk & Chothia, 1984) and has been refined in numer-

ous studies (e.g. Gerstein & Altman (1995)). This is done because the small distances between well-matched alpha carbon atoms have much less of an effect on the RMS than do the very large distances between poorly matched atoms. The untrimmed score of divergent protein domains is then concerned primarily with the poorly matched residues instead of the conserved core. Trimming alleviates this effect by restricting the RMS calculation to include only those residues believed to be in the conserved core. However, the degree of trimming is to some extent arbitrary, and this choice affects the baseline of the reported RMS scores. Here we considered only the better half (50%) of matched residues in a given pair of protein domains. Chothia & Lesk (1986) chose a somewhat different threshold. Figure 2(c) and (d) demonstrate the effect of trimming.

Analogous alignment similarity scores: Smith-Waterman score and structural comparison score

The dependence of the RMS separation on trimming method restricts its usefulness in comparing data. Likewise, there are many problems with using percent identity as a measure of sequence similarity. For instance, a match of non-identical but still similar residues (e.g. Arg *versus* Lys) scores the same as one between completely different residues (e.g. Arg *versus* Val), and gaps do not enter in the score calculation. Consequently, we now turn to alignment similarity scores, which eliminate some of the problems with traditional scores.

For sequence alignments, an alignment score is defined as the sum of the similarity matrix values for the alignment, minus the total gap penalty. This is sometimes called the Smith-Waterman score (Smith & Waterman, 1981). An analogous alignment score for structure is the structural comparison score, described by Levitt & Gerstein (1998). We will refer to these two similarity scores as S_{seq} and S_{str} , respectively. Note that they both increase for more similar pairs, whereas RMS increases for more divergent pairs. Specifically, S_{str} is the score maximized by the structural alignment program we used (Gerstein & Levitt, 1998). It can be calculated from any pair of aligned structures according to the function:

$$S_{\text{str}} = M \sum \left(\frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} - \frac{N_{\text{gap}}}{2} \right) \quad (1)$$

M and d_0 are constants, usually set to 10 and 5 Å, N_{gap} is the number of gaps in the alignment, d_i is the distance between each aligned pair of C^α atoms, and the sum is carried over all aligned pairs, i .

The main advantage of S_{str} over RMS in describing structural similarity is that the C^α to C^α distance, d_i , appears in the denominator of the calculation. This means that the smallest distances, corresponding to the best matches in the conserved core, are most significant in determining the score. Hence, the need for trimming is eliminated. S_{str} is also advantageous because it takes gaps into account and because of the fundamental analogy between this score and S_{seq} .

Figure 3(a) displays the relationship between structural and sequence similarity as expressed by S_{str} and S_{seq} . Figure 3(c) and (d) show calibration curves relating each of these scores back to approximate RMS separation and percent identity, respectively. Calibration curves help one get an intuitive feel for the degree of relationship in terms of the more traditional scores. Figure 3(b) adds a third axis, alignment length, and demonstrates that S_{str} depends greatly on this quantity. Although S_{str} and S_{seq} are "better" scores than RMS and percent sequence identity, the heavy dependence of both of these on length limits their usefulness in many situations. In other words, two pairs of similar domains with equal percent sequence identities but different lengths can have drastically different S_{seq} scores.

Probabilistic scores: P-values expressing the significance of sequence and structure similarity

Probabilistic scores can, to a great degree, overcome the length-dependence problems associated with the alignment scores. Probabilistic measures are advantageous because they express similarity not by an arbitrary "score" but by a statistical significance: the likelihood that such a similarity could be achieved by chance. This likelihood is also called the "P-value." We used calculations (described in detail in the legend to Figure 4) based on those given by Levitt & Gerstein (1998) to obtain P-values based directly on S_{str} and S_{seq} ; we refer to these calculated P-values as P_{str} and P_{seq} , respectively. For P_{seq} we could equally well have used the numbers from one of the popular sequence search programs (i.e. BLAST or FASTA) as all these values have been shown to be perfectly proportional to each other (Levitt & Gerstein, 1998; Brenner *et al.* 1998).

P_{seq} and P_{str} can be used to express the relationship between structure and sequence similarity on a more fundamental level. Figure 4(a) shows a log-log (base 10) plot of P_{str} against P_{seq} . Because it is log-log, trends can be visualized as straight lines. Two straight lines are necessary to fit the points well, with the discontinuous boundary between the lines located at the beginning of the twilight zone. The different slope of the line at low sequence similarity reveals that in the twilight zone there is a different relationship between the significance of structural similarity and that of sequence similarity. In particular, for domain pairs

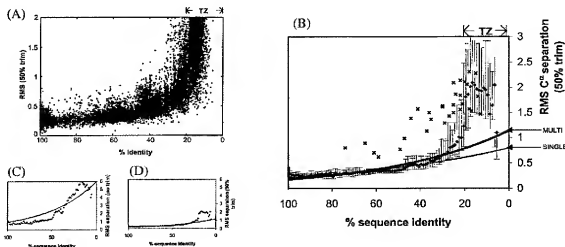


Figure 2. RMS as a function of percent identity. (a) A simple scatter plot of our pairs, relating RMS separation to percent sequence identity. This is similar to the presentation given by Chothia & Lesk (1986), but in this survey we looked at 30,000 pairs, 1000 times the number they compared. Outliers (pairs with RMS scores further than two standard deviations from the mean for their percent identity) are excluded from this graph; they represent domains that are very closely related with the exception of a conformational change. (b) A simplified graph with a number of fits to the data. For each percent identity bin we show the median RMS value, indicated by (♦) and the top and bottom quartile RMS values, indicated by the bars. Two fits are drawn through the median RMS values. The thin line, labeled SINGLE, is a simple exponential fit through the medians. It has the form:

$$R = 0.21e^{0.0132H}$$

where R is the RMS deviation after least-square fitting, H is the percent difference between the sequences (H for Hamming distance), and $H = 100\% - I$, where I is the percent sequence identity. The thick line, labeled MULTI, is a multigraph fit, which is described in the legend to Figure 4. The relation between RMS and percent identity according to this fit is expressed by the equation:

$$R = 0.18e^{0.0187H}$$

The twilight zone of sequence identity and below is labeled TZ. In this region, sequence similarity is not significant and not reliable for predicting structural similarity. This is why the median values in this area of the graph deviate significantly from the fits, which consider only data above 20% sequence identity. For reference we include the original data points from Chothia & Lesk's, 1986 paper (A.M. Lesk, personal communication), indicated by X. Their data follow the form:

$$R = 0.40e^{0.0187H}$$

The difference between the Chothia & Lesk trend and our relationship is due to the different trimming methods used in calculating the RMS score. Chothia and Lesk imposed a 3 Å cut-off in determining the conserved core residues; we defined the core as the better matching (in terms of Cα distances) half (50%) of the residue pairs. (c) and (d) The effect our trimming has on median RMS values. The RMS values in (c) are calculated from all the matched residues in each pair; the values in (d) are calculated from the better matching 50% of the residues.

in the twilight zone (according to the percent identity to P_{seq} calibration in Figure 4(b)), structural similarity is more significant than sequence similarity (having a smaller P -value or more negative $\log P$ -value). In contrast, for pairs with more than ~30% identity, the situation is reversed, with a given pair having more significant sequence similarity than structural similarity. One possible interpretation of this reversal is as follows. Structure is always more highly conserved than sequence, so usually a given amount of structural similarity is not as significant as a corresponding amount of sequence similarity. However, this is true only when meaningful sequence similarity

actually exists; thus, it does not apply in the twilight zone, where sequence similarity is by definition not significant. Note that all pairs in our comparison share at least the same fold, implying that they always have a significant amount of structural similarity.

In other words, for closely related sequences, differences in sequence similarity are more meaningful, whereas for highly diverged sequences that share the same fold, the differences in structural similarity are more significant.

Fitting two lines to the P_{str} versus P_{seq} graph suggests that the same might be done for other scoring schemes. It is possible to some degree to fit

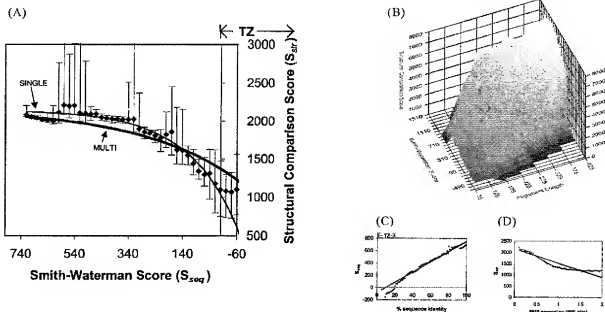


Figure 3. Similarity scores: structural comparison score as a function of Smith-Waterman score. Alignment similarity scores S_{str} and S_{seq} have certain advantages over RMS and percent identity scores for expressing the sequence-structure relation. S_{str} is calculated according to equation (1) in the text (Gerstein & Levitt, 1998; Levitt & Gerstein, 1998). S_{seq} is calculated using the BLOSUM50 matrix (Henikoff & Henikoff, 1992) with gap opening and extension penalties of -12 and -2 , respectively. (a) This is analogous to (b) in Figure 2. From the original 30,000 pairs we show the median S_{str} value for each S_{seq} bin, along with quartile bars above and below. Again the twilight zone and below is labeled TZ. The thin line, marked SINGLE, is a simple fit to the median S_{str} values in this graph; it has the form:

$$S_{str} = 2144 - 1106 \exp(-0.00544 S_{seq})$$

The thick fit, marked MULTI, is the multigraph fit, explained below. It follows the equation:

$$S_{str} = 2157 - 787 \exp(-0.0028 S_{seq})$$

The equations presented here provide an approximation of the observed trends; as (b) illustrates, they are nothing more than simple approximations. The main disadvantage of S_{str} as a measure of structural similarity is its heavy length dependency for pairs of structurally similar protein domains. (b) Surface plot of the median S_{str} as a function of S_{seq} and alignment length (the number of matched residue pairs). It is clear that the size of the aligned domains plays a major role in the resulting S_{str} , even though our fits do not take length into account. (c) and (d) Relate S_{seq} and S_{str} to the more familiar percent identity and RMS measures. The fits were used to convert between scoring schemes in constructing the multigraph fit. We derived the multigraph fit in order to create one set of equations and parameters that would relate sequence and structural similarity using either the percent identity and RMS scheme or the S_{seq} and S_{str} scheme, and allow translation between them. We simultaneously performed least-squares fits to the median values in four graphs: Figures 2(b) and 3(a) and the calibrations of S_{seq} to percent identity and S_{str} to RMS, (c) and (d), respectively. In all cases, we ignored data in and below the sequence identity twilight zone (labeled TZ). The parameters in (a) are dependent on the parameters in Figure 2(b) *via* the mentioned calibrations.

the traditional RMS *versus* percent identity graph (Figure 2) with two straight lines instead of an exponential curve. However, in this case, we opted for the more conventional presentation.

Class differences

The division of SCOP into classes based on secondary-structural composition allows easy investigation as to whether there are any deviations from the common similarity relationships on account of secondary-structure characteristics. Figure 5(a) reveals that secondary structural composition does not markedly affect the trends in sequence and structure similarities. This is consistent with the

data given by Wood & Pearson (1999). However, the larger average length of α/β domains compared with domains in the other classes results in a deviation in the length-dependent S_{str} (Figure 5(b)). The consistency among length-independent scores applies for certain individual folds as well. The immunoglobulin fold makes up an appreciable fraction of all the β -pairs (Figure 1(c)), yet the results are not affected if these pairs are left out.

Linking sequence and structure to function

Difficulties of functional comparison

There is a clear, well-characterized relationship between sequence and structure similarity, which

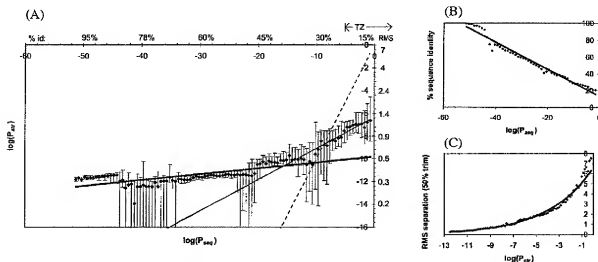


Figure 4. Probabilistic scores: P -values. P_{seq} and P_{str} are P -values calculated from S_{seq} and S_{str} according to the formalism given by Levitt & Gerstein (1998). Both quantities have the same overall functional form in terms of an extreme value distribution:

$$P = 1 - \exp(-\exp(-Z))$$

where P is either P_{seq} or P_{str} . For P_{seq} , $Z = S_{seq}/a - 2 \ln M - b/a$, where $a = 5.84$, $b = -26.3$, and M is the geometric mean of the lengths of the two sequences (i.e. $M^2 = nm$, where n and m are the two sequence lengths). For P_{str} , Z is a function of S_{str} and N , the number of matched residues: For $N < 120$:

$$Z = (S_{str} - c \ln^2 N - d \ln N - e)/(f \ln N + g)$$

For $N \geq 120$:

$$Z = (S_{str} - a \ln N - b)/(f \ln 120 + g)$$

At $N = 120$, continuity implies that:

$$a \ln 120 + b = c \ln^2 120 + d \ln 120 + e \quad \text{and} \quad a = 2c \ln 120 + d$$

This, in turn, allows the calculation of the constants:

$$a = 171.8, \quad b = -419.4, \quad c = 18.4, \quad d = -4.50, \quad e = 2.64, \quad f = 21.4, \quad g = -37.5$$

(a) of this Figure is analogous to Figures 3(a) and 2(b), with the exception of the fits. It is a log-log (base 10) plot relating P_{seq} and P_{str} . We show the median $\log(P_{str})$ value for each $\log(P_{seq})$ bin, along with quartile bars above and below. We have added approximate percent identity and RMS values to the x and y axes to aid interpretation of the graph in terms of more familiar scores. The values were calculated using the calibration curves in (b) and (c). The straight-line nature of the log-log plot reveals distinct relations inside and outside the twilight zone, labeled TZ. (The area of percent identity below the twilight zone does not appear in P_{seq} graphs, there is no significance for such low sequence similarity; thus all data points in that zone appear at $P_{seq} = 1$ or $\log(P_{seq}) = 0$.) The thick line in the figure is fit to the median P_{str} values for P_{seq} values outside the twilight zone; its equation is:

$$P_{str} = 10^{-10} P_{seq}^{0.05}$$

The thin line is fit to the data inside the twilight zone; it follows the relation:

$$P_{str} = 10^{-6} P_{seq}^{0.274}$$

For reference we include the dotted line, representing the function $P_{str} = P_{seq}$ where sequence and structural similarity are equally significant. See the text for a discussion of how the two trends might be interpreted with respect to this line.

can be used to transfer precisely structural annotation based on the degree of sequence homology. In genome analysis, however, one is usually more interested in finding a functional annotation for an open reading frame based on similarity to well-

known proteins; yet the sequence-function and structure-function relationships have not been as explicitly characterized. The fundamental obstacle to extending this and similar investigations to deal with function is the absence of a clear measure of

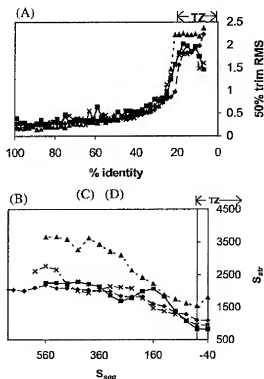


Figure 5. SCOP class differences. Previously it has been observed that secondary structural composition does not cause deviations from the trends in structure and sequence similarity (Flores *et al.* 1993). To test this observation we looked at the scores divided by SCOP class. The following legend applies to the graphs: (—■—), all alpha; (---◆---), all beta; (---▲---), alpha/beta; (- - - × - -), alpha + beta. (a) Median RMS values for each percent identity bin. The traditional scores reveal no dependency on class. However, in (b) α/β pairs consistently score higher S_{tr} scores than pairs in other classes. This is a consequence of the dependence of S_{tr} on length; domains in the α/β class are longer, on average, than in the other classes.

functional similarity. Although we were able to present three different quantitative measures of structural relatedness, an analogous situation for function does not exist. How can one express quantitatively the degree of similarity between a triosephosphate isomerase and a glucose-6-phosphate isomerase? How do they compare to trp repressor?

The absence of a clear measure of functional similarity is not the only obstacle in transferring the functional annotations between proteins with different degrees of homology. The definition of function itself is often vague. More specifically, at present there is an absence of such important information as a standardized vocabulary for protein functional annotations with an associated numbering scheme, descriptions of monomer functions of subunits of multisubunit proteins and hierarchical functional assignments for proteins with multiple

functions. As a consequence of these difficulties there is no functional equivalent to the hierarchical fold classification for domains in PDB.

As signs of progress in this direction, several functional classifications have been developed to date. One is the ENZYME system developed by the Enzyme Commission (EC) to classify enzymes by reaction type (Webb, 1992). This system has the advantage that it is "universal," applicable to proteins in many different organisms, and is in wide use. However, it also has several drawbacks. First of all, it does not consider catalytic reaction mechanisms (Riley, 1998a), often ignoring obvious similarities. Second, it presumes a 1:1:1 relationship between gene, protein and reaction, although this is often not the case (an enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function). Perhaps the most significant drawback of the EC classification is that it applies to only enzymes.

A number of more comprehensive schemes have been developed, which classify non-enzymes as well as enzymes. Most of these focus on individual organisms. Several such schemes exist, for instance, GenProtEC/EcoCyc for *E. coli* (Karp *et al.*, 1998b; Riley & Labedan, 1996; Riley, 1998b), MIPS for yeast (Mewes *et al.*, 1998), Ashburner's functional classification for *Drosophila*, which is connected to FLYBASE (Ashburner & Drysdale, 1994), and ECAD for human ESTs (Adams *et al.*, 1995). These classifications possess some advantages. They have additional levels of hierarchy that help present a more comprehensive picture of genotype-phenotype relationships. On the other hand, these classifications still leave much room for improvement. For example, there is no standardized vocabulary to allow for keyword searches among multiple databases and across organisms, and there are inconsistencies in category numbering style.

Finally, there has been some promising work going beyond the ENZYME and organism-focused classifications. There has been progress on completely automated functional classification (des Jardins *et al.*, 1997; Tamames *et al.*, 1997), which has the potential for putting function assignments on a more objective basis. There are a number of databases synthesizing the various enzyme functions into coherent pathways and systems (e.g. KEGG and WIT, Ogata *et al.*, 1999; Selkov *et al.*, 1998). There also have been some very recent attempts to develop cross-species classifications of non-enzyme functions in the framework of the Gene Ontology Project (GO, geneontology.org). GO is a joint project between FlyBase, the Saccharomyces Genome Database and Mouse Genome Informatics, attempting to merge the fly, yeast and mouse functional classification schemes. However, a truly universal system for classifying all protein functions in all organisms within the same framework remains quite a challenge because of the

sheer diversity of organisms and distinct protein functions.

Our simple functional classification of SCOP domains: FLY+ ENZYME

Given the discussed limitations, we constructed a simple functional classification for the SCOP domains included in our comparison; our classification is based on a merger of two of the existing functional annotations and a cross-referencing of subsets of this combination with some of the organism-specific schemes. First, we used pairwise comparison to cross-reference the PDB domains against the Swissprot database (Bairoch & Apweiler, 1998), as described by Hegyi & Gerstein (1999). We chose to assign protein functions according to Swissprot because it provides more comprehensive functional annotations than SCOP.

We were initially able to divide all entries into enzymes and non-enzymes, a division that represents the highest level of functional difference in our classification scheme (Figure 6). For the enzyme category, we transferred EC (Webb, 1992) numbers to those SCOP domains with a one-to-one match to a Swissprot enzyme. Only one-to-one matching entries could be considered because Swissprot assigns ENZYME numbers to entire proteins, whereas SCOP is a domain-based classification; therefore we could be confident about the classification of only those domains which map to an entire Swissprot entry.

In the absence of an EC-type classification for non-enzymes, we assigned functions to non-enzymatic SCOP domains according to Ashburner's original classification of *Drosophila* protein functions. This classification is derived from a controlled vocabulary of fly terms. It is available on the web and loosely connected with the FLYBASE database (Ashburner & Drysdale, 1994). For clarity, we precisely describe the specific files and version (1.55, 1997) of the classification that we used in the caption to Figure 6, and we will hereafter refer to these data files as constituting the original FLY classification.

The FLY classification is a dynamic object, changing as more is learned about the fly and other organisms. This is particularly true of late with the imminent completion of the *Drosophila* genome. In fact, since the completion of our analysis, the FLY classification has been superseded by the new GO classification (see above).

The hierarchical structure of the FLY classification makes it well suited for classifying non-enzymatic SCOP entries in a manner comparable to the ENZYME assignments for the enzymes. Another advantage of this classification is that it is more compatible with the makeup of the PDB than the *E. coli* and yeast classifications, as *Drosophila* is a multi-cellular organism, and many of the known structures come from animals. We were able to use the original FLY classification as a framework to

which we added functional categories and individual proteins. For instance, we added "Hemoglobin" to the "Physiological Processes - Respiration" category. Another example is the "Physiological processes - Immunity" category (Figure 6(b)), to which we added immune system proteins. Many of the additions would not be necessary in the context of the new cross-species GO system. We also modified slightly the numbering scheme in the original FLY classification in order to assign a unique hierarchical number to each protein domain (Figure 6(b)). We will refer to our augmented FLY classification as the FLY+ scheme, and our merged scheme as the FLY+ ENZYME classification.

As discussed earlier, the universal functional classification of proteins is very challenging and may not be possible with the current level of knowledge about genes, proteins and genomes. Consequently, the FLY+ ENZYME classification of SCOP proteins is somewhat incomplete and inconsistent and retains many of the limitations of its components (Hegyi & Gerstein, 1999; Riley, 1998a). It is not yet broad enough to include many plant, virus and bacterial proteins. Nevertheless, it was sufficient for our analysis, as we were able to classify a very large number of the total 30,000 pairs.

Determining functional similarity

Using our compound functional classification, we were able to assign a level of functional similarity to each domain pair. According to our scheme, a pair can have no functional similarity (an enzyme paired with a non-enzyme) or it can have one of three levels of similarity:

- (i) General similarity. Both domains are enzymes or both are non-enzymes.
- (ii) Same functional class. Both domains share the first component of their ENZYME or FLY+ numbers, e.g. 1.1.1.1 alcohol dehydrogenase and 1.3.1.1 cortisone beta-reductase (for enzymes), or 3.3.2.1.2 calcyclin and 3.6.3.2.1 calmodulin (for non-enzymes).
- (iii) Same precise function. Both domains share three components of their ENZYME or FLY+ number, e.g. 1.1.1.1 alcohol dehydrogenase and 1.1.1.3 homoserine dehydrogenase (for enzymes) or 1.2.9.1.1.1 Arc repressor and 1.2.9.1.1.1 C-jun (for non-enzymes; both are transcription factors). A pair that shares precise function must also, by definition, share functional class and general similarity.

Based on those assignments we calculated the percentage of total pairs at a given level of sequence or structural similarity possessing each level of functional similarity. The results appear in Figure 7.

Sequence and function

The relation between sequence similarity and functional similarity behaves as one might expect, with sigmoidal curves that drop off sharply at particular conservation thresholds, and with the three levels of functional similarity (precise function, functional class and general similarity) having progressively lower thresholds. Figure 7(a) shows that precise function is not conserved below 30–40% sequence identity, whereas functional class is conserved for sequence identities as low as 20–25%. Below 20%, general similarity is no longer conserved; among pairs of approximately 7% sequence identity, about 40% are enzymes paired with non-enzymes. It is important to note that in all the pairs considered here, the domains share the same fold. Functional similarity at low percent identities (e.g. 7%) would be much less for all possible pairs of domains rather than just for those with the same fold. It is also important to remember that our thresholds for functional conservation are statistical averages over many sequences; one will, of course, be able to find individual cases that diverge more or less rapidly.

There are differences between the functional conservation thresholds of enzymes and non-enzymes, with enzymes appearing to more highly conserve precise function than non-enzymes, but non-enzymes conserving functional class more highly than enzymes. This may reflect that in our classification, the non-enzyme functional classes are broader and hence easier to conserve than those of the enzymes, while the non-enzymatic precise functions are more specific.

When P_{seq} is used as the measure of sequence similarity (Figure 7(b)) the results look somewhat different, it appears that functional class is conserved for the entire range of sequence similarities. In this case, percent identity is actually more discriminating than P_{seq} because functional class diverges only at sequence similarities that are low enough that they have little or no statistical significance, i.e. for P_{seq} the divergence is compressed near the vertical axis of the graph.

Structure and function

The relation between similarity in structure and function is somewhat less straightforward than that between similarity in sequence and function. Figure 7(c) shows the relationship between RMS and functional similarity. Broadly, it appears similar to that for percent identity and functional similarity; however, the thresholds for conservation of the various types of functional similarity are less sharp.

RMS is more revealing with respect to functional similarity than the non-traditional structural scores, S_{str} and P_{str} . (Data for S_{str} and P_{str} are not shown but are available from the website.) The reason is that, while very structurally similar pairs all have RMS scores clustered between 0 and 0.5 Å, S_{str} has

a large range of scores for similar pairs due to the length dependency, and P_{str} does not have any limit for maximum similarity. The wide range of possible S_{str} and P_{str} scores for similar structures tends to blur the broad sigmoid curves so much so that they are no longer apparent.

Alternative functional classifications: MIPS and GenProtEC

To get some perspective on the degree to which our results reflected the particularities of our combined FLY + ENZYME classification, we decided to try the same comparisons based on the well-known functional classifications for yeast and *E. coli*, MIPS and GenProtEC (Mewes *et al.*, 1998; Riley & Labedan, 1996; Riley, 1998b). These classifications have the advantage that they integrate enzyme and non-enzyme functions from the start and are widely used. However, as they are only applicable to individual organisms, we could only use them to classify a considerably smaller subset of the known structures than the compound FLY + ENZYME system.

The specific way we used the MIPS and GenProtEC classifications to assign function to structures and to calculate functional similarities is described in the legend to Figure 7. Our results in terms of functional conservation (precise and class) at various levels of percent identity are shown in Figure 7(d). We observe the same general relationships as we did for our FLY + ENZYME scheme. That is, the functional conservation curves have a sigmoidal shape and have cut-offs for precise functional similarity after 40% and for functional class similarity at lower values. However, because the MIPS and GenProtEC classifications are restricted to individual organisms, each curve represents considerably fewer data points than do the curves based on the FLY + ENZYME scheme; this required us to “bin” the MIPS and GenProtEC curves in a somewhat coarser fashion.

Discussion and Conclusion

Here, we assessed the transfer of functional and structural annotation by analyzing the relationships between similarity in sequence, structure and function. The ~30,000 protein domain pairs of varying levels of similarity (at least the same fold) that we constructed out of the SCOP classification show quantitative sequence-structure relationships consistent with previous research. The exponential relationship is consistent across the secondary-structural classes and holds for newer probabilistic scoring methods.

The sequence-function and structure-function relationships have not been studied as precisely due to the lack of a robust functional classification and measure of functional similarity. To overcome

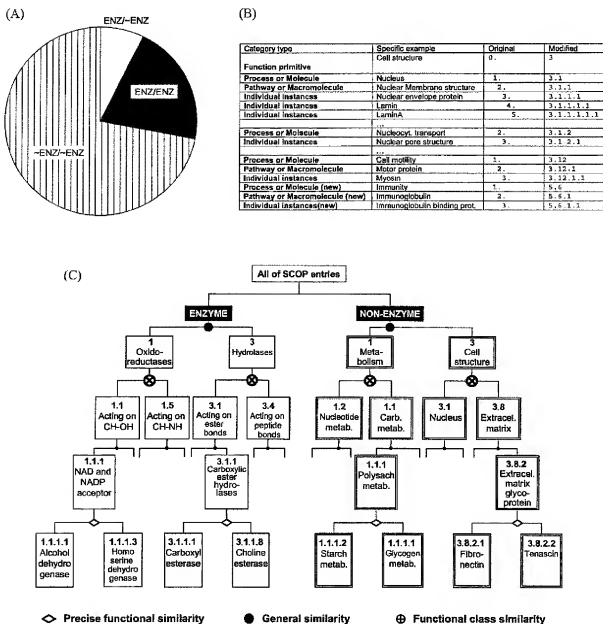


Figure 6. Functional classification of enzymes and non-enzymes. (a) Divides the pairs by general function. There are three categories of pairs: (i) enzymes paired with non-enzymes (no general functional similarity), labeled ENZ/~ENZ; (ii) enzymes paired with enzymes (same general function), labeled ENZ/ENZ; and (iii) non-enzymes paired with non-enzymes (same general function). Pairs for which one or both domains could not be identified as enzyme or non-enzyme are not included in this chart. Enzymes are classified according to the EC system (Webb, 1992). The first component of the number represents the nature of reaction and is called class. There are six classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The next level is subclass. It refers to the chemical groups on which the enzyme acts. For example, the first class, oxidoreductases, has 19 subclasses that are arranged according to the donor group that undergoes oxidation (CH-OH, aldehyde or oxo group, CH-CH group, etc). For another group of enzymes (hydrolases) subclass is determined by the nature of the bond: ester bond, peptide bond, etc. The next level is sub-subclass. For oxidoreductases this indicates the acceptor group: NAD(+) and NADP(+), or cytochrome; for hydrolases the sub-subclass represents the nature of substrate (carboxylic ester hydrolases, thiolester hydrolases, etc.). The fourth level represents a unique number for each individual enzyme, for example, 1.1.1.1: alcohol dehydrogenase. (b) Shows how we adapted the functional classification of *Drosophila* gene products developed by M. Ashburner. This classification is loosely connected with FLYBASE (Ashburner & Drysdale, 1994). We used version 1.55 (4 August 1997) that was available from Ashburner's website:

<http://www.ebi.ac.uk/~ashburn>

The specific files that we used were taken from the ftp directory:

<ftp.ebi.ac.uk/databases/edgpc/misc/ashburner>

this we constructed our own classification by merging and extending the ENZYME and FLY schemes and assigning levels of functional similarity. Our measures of functional similarity provide curves relating function to sequence and structure; when relating functional conservation to sequence divergence, we find distinct thresholds at ~40% for precise function and ~25% for functional class.

One of the interesting results that emerges from this is that percent identity is more useful for quantifying functional divergence than the newer probabilistic scores. In general, modern probabilistic scores, such as P_{seq} are better at discriminating amongst highly diverged sequences (near the twilight zone) than percent identity, since they better take into account gaps and conservative substitutions (of similar amino acids). However, for very similar pairs of sequences, percent identity is a simpler and more direct measure of divergence (essentially a Hamming distance). Since divergence in precise function takes place before that in structure (well before the twilight zone), it is quite reasonable that percent identity is more successful at measuring the former than the latter and that

the converse is true for the probabilistic scores. In other words, percent identity is better calibrated for discriminating amongst very close, significant relationships and P_{seq} for more distant ones.

Practical implications

The sequence-structure and sequence-function relationships described here provide practical information for genome annotation in terms of folds and functions. Table 1 summarizes the relative advantages of the different scoring methods we used. Using the trends in sequence and structure similarity, one can assess the degree to which structural annotation can be transferred between sequences at a given level of sequence similarity. The sequence and function similarity thresholds potentially establish minimum requirements of sequence similarity for reliable function prediction. Note that because the protein domain pairs considered here all share the same fold, the numbers for all possible pairs will differ in the region of very little sequence identity, in which the sequence similarity is not enough to indicate the same fold.

We refer to these as constituting the original FLY classification. Recently, the FLY classification has been superseded by the GO (Gene Ontology) Project classification, which merges fly, mouse and yeast annotation. Files related to the GO classification are available from www.geneontology.org. In the original FLY classification all members of the highest level are labeled 0, representatives of the next level are labeled 1, and all lower levels are labeled 2 through to 9. We changed the numbering scheme so that it will reflect the hierarchical nature of the classification. This Figure illustrates sections of the original and modified classification. The top level in the FLY classification scheme is called "Function primitive" (level 0) and includes five classes: "Metabolism," "Intracellular protein traffic," "Cell structure," "Developmental process," "Physiological process," and "Behavior." The next level after "Function primitive" is "Process" or "Molecule" (level 1 in Ashburner's classification). For "Function primitive - Metabolism" the processes are "Carbohydrate metabolism," "Nucleotides and nucleic acids metabolism," etc. For "Function primitive - Cell Structure" the "Process" can be "Nucleus," "Mitochondrion," "Membrane," etc. The next level is "Pathway" or "Macromolecule" (level 2 in the original classification). "Pathway" can include "Metabolic pathway," "Signaling pathway," or "Developmental pathway." The "Macromolecule" category includes "Protein" and "Nucleic Acid." We added categories to the original classification in order to classify some mammalian proteins that are widely represented in SCOP but are absent from the original FLY scheme. These categories include immune system proteins (labeled "new" in (b) and respiratory proteins such as hemoglobin and myoglobin that we added to "Function primitive - Physiological process - Respiration"). We call our adaptation of the original FLY scheme, FLY+. Further information on this adaptation is available at:

<http://bioinfo.mbb.yale.edu/align/func>

(c) The overall hierarchy of our final scheme and identification of the different levels of similarity. If two proteins are both enzymes or both non-enzymes, then they possess general functional similarity. If they share the first component of their classification numbers, then they are in the same functional class. If they share the first three components of their enzyme numbers (or the equivalent for non-enzyme numbers, depending on category) then they have the same precise function. A significant difference between the two main branches of the hierarchy is that the levels of the ENZYME classification do not correspond exactly to those in the FLY+ system because the fly classification is more extensive than the enzyme classification. For instance, the FLY classification takes into account aspects of cellular (cytoskeleton, metabolic pathways, etc.) and phenotypic function (morphology, physiology, behavior) that are absent from the ENZYME scheme. This makes our classification of SCOP proteins somewhat unbalanced, as non-enzymes have much broader and more loosely defined functional classes. As a consequence, while each enzyme is assigned a four-component number, the length of a non-enzyme number varies, depending on the functional category to which it belongs. For example, myosin is assigned a number that happens to have the same length as EC numbers: 3.12.1.1. However, transcription factors are numbered 1.12.9.1.1.1. We took into account this varying hierarchy depth in deciding how many components are necessary to identify precise function in each category. Note that what we mean by domains having the same precise function is not the same as the domains coming from the same essential protein.

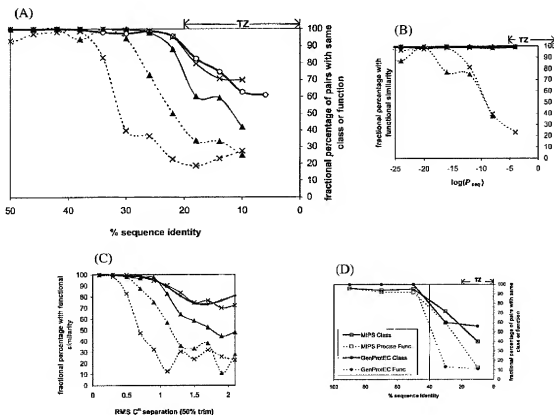


Figure 7. Linking sequence, structure and function. We express functional similarity as the fractional percentage of pairs at a given level of sequence/structural similarity for which the paired domains share a precise function, functional class, or general similarity (according to our classification, see Figure 6). The following legend applies to (a) through (c): (—○—), general similarity; (—×—), non-enzymes with same functional class; (—△—), enzymes with same functional class; (---×---), non-enzymes with same precise function; and (---△---), enzymes with the same precise function. (a) Relates functional similarity to sequence similarity in terms of percent identity. The functional similarity appears as a sharp sigmoid, with distinct thresholds of divergence for precise function, functional class, and general similarity. Enzymes are paired with non-enzymes only at very low percent identity, in and below the twilight zone (labeled TZ). At slightly higher sequence identity, pairs diverge with respect to functional class, and beyond 40% identity with respect to precise function. Note that 50-100% identity is not shown because almost all domains that are that similar share function with their counterparts. (b) Shows the same data using P_{seq} as the measure of sequence similarity. Only the divergence in precise function is visible because there is such little significance for the low sequence similarity at which functional class and general similarity diverge, all data points in that region appear near $P_{seq} = 1$ or $\log(P_{seq}) = 0$ (the y-axis). (c) Illustrates that the structure-function relation is not as clearly defined as that for sequence and function. Functional similarity expressed in terms of RMS separation appears as a broad sigmoid curve; there are thresholds of divergence for precise function, but the divergences in functional class and general similarity are more gradual. The thresholds are apparent only because RMS clusters the most structurally similar pairs between scores of 0 and 0.5 Å. For this reason, RMS is better at discerning functional similarity than S_{50} and P_{seq} , which do not cluster the most similar pairs around a set limit. (d) Shows the same relationships (functional conservation versus percent identity) as in (a), except that for this graph functional similarity is determined in terms of the MIPS (Mewes *et al.*, 1998) and GenProtEC (Riley, 1998b) classifications rather than the FLY + ENZYME scheme. The legend appears as the inset on the graph. We assigned MIPS and GenProtEC classifications to SCOP domains based on sequence comparisons to classified yeast and *E. coli* open reading frames (ORFs), respectively. The SCOP domain most closely matching each ORF classified in MIPS or GenProtEC was assigned the corresponding MIPS or GenProtEC function number. Only matches of 80% sequence identity or greater were considered. We used this SCOP domain as a functional representative; when determining functional similarity, we assigned to SCOP domains with no MIPS or GenProtEC functional designation the function of the closest representative with at least 85% sequence identity, if one existed. GenProtEC functional identifiers are three-component numbers. We consider a pair of domains sharing the first component of their functional designation to be in the same functional class. Domains that share all three components are said to have the same precise function. For MIPS the functional designation is not as straightforward, as one ORF can be assigned multiple functions. Therefore we consider domains which have at least one function in common to share functional class. Domains with all functions in common, the same combination of identifiers, share precise function. Because MIPS and GenProtEC each classify the proteins of a single organism, yeast and *E. coli*, respectively, these classifications can determine the functional similarities of only a small fraction of all our SCOP domain pairs. The data based on these classifications, appearing in (d), are therefore very sparse compared to the data in (a)-(c). Despite the coarseness of the data, functional similarity based on the MIPS and GenProtEC classifications follows the same general relation to sequence similarity as does functional similarity based on the more comprehensive FLY + ENZYME scheme. Vertical line indicates an approximate threshold of functional divergence at 40% identity.

Table 1. Summary of scoring methods

	Sequence similarity	Structural similarity	Features	Limitations
Traditional scores	Per cent sequence identity	RMS C $^{\alpha}$ separation	Well understood, in use; percent identity better for looking at functional similarity	RMS depends most highly on worst matches, requiring arbitrary trimming; percent identity is insensitive to gaps and conservative substitutions
Alignment similarity scores	S_{seq}	S_{str}	Analogous similarity scores, S_{str} depends most highly on best matches	Dependence on alignment length
Modern probabilistic scores	P_{seq}	P_{str}	Statistical significance, unified framework for different comparisons	Not as familiar as RMS and percent identity

The Table lists the schemes presented here for characterizing the sequence-structure relationship, along with their relative advantages and disadvantages.

Practically, then, when one searches an uncharacterized open reading frame against known structures, if the open reading frame matches a structure with a good *e*-value or percent identity, then the curves presented here can be used to check how the functional and detailed structure annotation will transfer. For example, if an unknown open reading frame matches a PDB structure with an *e*-value of 0.001 and a percent identity of 30%, then one can be assured that it has the same fold (Brenner *et al.*, 1998) and according to our analysis it has a two-thirds chance of having the same exact function. Furthermore, it has a ~99% chance of having the same functional class and its structure probably diverges from the known structure by a trimmed RMS of less than 0.7 Å.

Future directions

There are a number of directions in which we might extend this analysis. With respect to the sequence-structure relation, we can reduce the overrepresentation of the immunoglobulins and improve the calculation of P_{str} (by redoing the fit to the extreme value distribution reported by Levitt & Gerstein (1998) to eliminate residual length-dependency).

In the functional realm, we can investigate if and how the sequence-function and structure-function relationships vary for different categories of proteins. For example, although we found consistency of the sequence-structure relationship among secondary structural classes, Hegyi & Gerstein (1999) found that the distribution of enzymes and non-enzymes varies with secondary structural class. A related issue is that of conformational changes. It is conceivable that among domains with very similar sequences but structures that differ by a conformational change, function is less conserved than it is among similar sequences with more similar structures.

Perhaps the most important direction in which to further this work is the augmentation of the functional classification. With the growing

amount of fully sequenced genomes there is a need for the development of a comprehensive system for functionally classifying proteins, a complete classification for the entire universe of protein functions. It will be a difficult process, as many existing organism-specific classifications will have to be merged, but the end result will have the advantage of not being biased towards any one organism. Such a universal classification will allow much more reliable transfer of functional annotation.

Acknowledgments

We thank A. Lesk for helpful conversations and supplying us with reference data for Figure 2, S. Brenner for providing carefully curated SCOP domain sequences, and H. Hegyi, W. Krebs and V. Alexandrov for assistance with the sequence comparisons, development of the FLY + ENZYME scheme, and design of the web database. M.G. thanks the Keck and Donaghy foundations for financial support.

References

- Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.
- Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., Venter, J. C., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3-174.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tools. *J. Mol. Biol.* **215**, 403-410.
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation

- of protein database search programs. *Nucl. Acids Res.* 25, 3389-3402.
- Andrade, M. A. & Sander, C. (1997). Bioinformatics from genome data to biological knowledge. *Curr. Opin. Biotech.* 8, 675-683.
- Ashburner, M. & Drysdale, R. (1994). Flybase: the *Drosophila* genetic database. *Development*, 120, 2077-2079.
- Attwood, T. K., Flower, D. R., Lewis, A. F., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N. & Wright, W. (1999). PRINTS prepares for the new millennium. *Nucl. Acids Res.* 27, 220-225.
- Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* 26, 38-42.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.
- Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* 6, 366-376.
- Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* 4, 393-403.
- Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707-725.
- Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* 15, 132-133.
- Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.* 266, 635-643.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, 95, 6073-6078.
- Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5, 236-244.
- Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5, 823-826.
- Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* 52, 399-405.
- des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*, 5, 92-99.
- Doolittle, R. F. (1987). *Of Urfs and Orfs*, University Science Books, Mill Valley, CA, USA.
- Enright, A. J., Iliopoulos, I., Kyriopoulos, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402, 86-90.
- Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence to structure-to-function paradigm with application to glutaredoxins/thioredoxins and T₁ ribonucleases. *J. Mol. Biol.* 281, 949-968.
- Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703-711.
- Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar domain pairs. *Protein Sci.* 2, 1811-1826.
- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Venter, J. C., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270, 397-403.
- Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren, E., Hardham, J. M., McLeod, M. P., Salzberg, S., et al. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, 281, 375-388.
- Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* 274, 562-576.
- Gerstein, M. (1998a). Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, 14, 707-714.
- Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* 33, 518-534.
- Gerstein, M. (1998c). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des.* 3, 497-512.
- Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* 251, 161-175.
- Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* 22, 277-304.
- Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *ISMB*, 4, 59-67.
- Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* 7, 443-456.
- Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288, 147-164.
- Heinikoff, S. & Heinikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915-10919.
- Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* 25, 236-239.
- Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87, 2264-2268.
- Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA*, 90, 5873-5877.
- Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* 20, 175-203.

- Karp, P. D. (1996). A protocol for maintaining multibase referential integrity. *Pac. Symp. Biocomput.* 438-445.
- Karp, P. (1998a). What we do not know about sequence analysis and sequence databases. *Bioinformatics*, 14, 753-754.
- Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998b). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* 26, 50-53.
- Karp, P. D., Ouzounis, C. & Paley, S. M. (1996b). Hincyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. *ISMB*, 4, 116-124.
- Lesk, A. M. & Chothia, C. (1984). Mechanisms of domain closure in proteins. *J. Mol. Biol.* 174, 175-191.
- Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, 95, 5913-5920.
- Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucl. Acids Res.* 26, 33-37.
- Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins: Struct. Funct. Genet.* 1, 2-6.
- Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
- Myers, E. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* 4, 11-17.
- Needleman, S. B. & Wunsch, C. D. (1971). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of genes and genomes. *Nucl. Acids Res.* 27, 29-34.
- Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* 273, 349-354.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284, 1201-1210.
- Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* 266, 227-259.
- Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276, 71-84.
- Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, 85, 2444-2448.
- Riley, M. (1998a). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* 8, 388-392.
- Riley, M. (1998b). Genes and proteins of *Escherichia coli* K-12. *Nucl. Acids Res.* 26, 54.
- Riley, M. & Labeledan, B. (1996). *E. coli* gene products: physiological functions and common ancestries. In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Neidhardt, F., Curtiss, R., III, Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umberger, H. E., eds), 2nd edn, pp. 2118-2202. ASM Press, Washington, DC.
- Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* 12, 85-94.
- Russell, R. B. & Barton, G. J. (1994). Structural features can be conserved in proteins with similar folds. *J. Mol. Biol.* 244, 332-350.
- Russell, R. B., Sagi, M. A. S., Sayle, R. A., Bates, P. A. & Sternberg, M. J. E. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* 269, 423-439.
- Russell, R. B., Sasienski, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds - binding site similarity in the absence of homology. *J. Mol. Biol.* 282, 903-918.
- Salamov, A. A., Suwa, M., Orenko, C. A. & Swindells, M. B. (1999). Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* 12, 95-100.
- Selkov, E., Jr, Grechkin, Y., Mikhailova, N. & Selkov, E. (1998). MPW: the metabolic pathways database. *Nucl. Acids Res.* 26, 43-45.
- Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-198.
- Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* 9, 368-373.
- Tamames, J., Casari, C., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* 44, 66-73.
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278, 631-637.
- Webb, E. C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York.
- Wood, T. C. & Pearson, W. R. (1999). Evolution of protein sequences and structures. *J. Mol. Biol.* 291, 977-995.
- Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res.* 26, 3986-3990.

Edited by F. E. Cohen


(Received 2 September 1999; received in revised form 5 January 2000; accepted 6 January 2000)

BLAST Basic Local Alignment Search Tool

Return to current design Edit and Resubmit Save Search Strategies Formatting options Download

Blast 2 sequences

SCS0009 v. PREF-1

Results for: Icd16501 None(352aa) 

Your BLAST job specified more than one input sequence. This box lets you choose which input sequence to show BLAST results for.

Query ID

Icd16501

Description

None

Molecule type

amino acid

Query Length

352

Subject ID

16502

Description

None

Molecule type

amino acid

Subject Length

385

Program

BLASTP 2.2.18+ Citation

Reference

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference: [compositional score matrix adjustment](#)

Stephen F. Altschul, John C. Vothton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", *FEBS J.* 272:5101-5109.

Other reports: [Search Summary](#) [Taxonomy reports](#)

Search Parameters

Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Threshold	11
Composition-based stats	2
Genetic Code	1
Window Size	40

Karlin-Altschul statistics

Params	Gapped	Ungapped
Lambda	0.267	0.325778
K	0.541	0.143002
H	0.14	0.515221

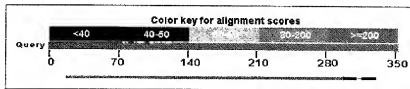
Results Statistics

Effective search space:114310

[Graphic Summary](#)

Distribution of 3 Blast Hits on the Query Sequence

An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple alignments on the same database sequence are connected by a striped line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments. *New*. This graphic is an overview of database sequences aligned to the query sequence. Alignments are color-coded by score, within one of five score ranges. Multiple alignments on the same database sequence are connected by a dashed line. Mousing over an alignment shows the alignment definition and score in the box at the top. Clicking an alignment displays the alignment detail.



[Dot Matrix View](#) **Plot of lcl|16501 vs 16502 [?]**

This dot matrix view shows regions of similarity based upon the BLAST results. The query sequence is represented on the X-axis and the numbers represent the bases/residues of the query. The subject is represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Plus strand and protein matches are slanted from the bottom left to the upper right corner; minus strand matches are slanted from the upper left to the lower right. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

**Descriptions**

Legend for links to other resources: [U](#) UniGene [G](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Sequences producing significant alignments:

(Click headers to sort columns)

	172	222	87%	2e-47
16502	unnamed protein product			

Alignments Select All Get selected sequences Distance tree of results

>cl116502 unnamed protein product
Length=385

Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

Score = 172 bits (435), Expect = 2e-47, Method: Compositional matrix adjust.
Identities = 111/319 (34%), Positives = 148/319 (46%), Gaps = 45/319 (14%)

Query 19 AFQGFVRADDCSHCDLHGCARPDGSCRCDFPQNEGLHCRCVMPGQCHGTCHQFMOCI 78
A G +C CD +G C D CRC GWEG C+CV PGC +G C -PWQCI
Sbjct 16 AFHSTYTGRECDPFCDFQYGTCEADNYCRCHVGWESPLCDKCVTARGCVNGVCKEPWQCI 75

Query 79 CHSGWAGKFCDK-----GFHGRDCERKAGHCEQAG 108
C GW GKFC+ GF G+DC+ KAGFC G
Sbjct 76 CRQGDGKPCIDIVRACTSTFCANNGTCVDEKGVQVCSCTPGFSKDCCKAGAPGVING 135

Query 109 SPCKNGGOCODDGFALNFTFCCLMGFVSARGEV--NVDDCLMRPCANGATCLDGINRES 166
SPC+HGG C DD+G S + +C C GF G CE+ + C PC N C D P
Sbjct 136 SPQCHGGACVDDGGQASHASCLCPGFSGMFCRIVAATNSCTPNPCENDGVCTDGGDFR 195

Query 167 CLCPGSGFAGRFCTINLDDCASRPCORGARCRVH-DFDCLCPSGYGGKTCCLWL-PVPD 224
C CP GF + C+ +CAS PCO G C F+CLC + G TC P
Sbjct 196 CRCPAGEVONTCSREVSNCASGPDQNGGTCLQHTQVSPCLCKEFPNGFTCAKKGASPV 255

Query 225 PFTTVDTPLGFTSAVV-----VPATGPAPHSAGAGLLRISVKEVVRROEAGLGEPSLVAL 279
T + + G T+ + +P P N PL+SVKE + + L E + +
Sbjct 256 QVTHLPSGYGLTYRLTPGVHLEFVQVQPEQH-----ILKVSMEK-LNKSTELLTEGGAICF 309

Query 280 VVFGALTAALVATVLLTL 298
+ G LT+ +VL TV +
Sbjct 310 TILGLVLSLVVLTVAIVF 328

Score = 36.2 bits (82), Expect = 2e-06, Method: Compositional matrix adjust.
Identities = 48/199 (24%), Positives = 71/199 (35%), Gaps = 51/199 (25%)

Query 117 CQDDCGF-AIAFTFCRLVGFVARGEVNVDDCLMRPCANGATCLGINR--FSLCLPEGF 173
C GF + CRC VG+ G C D C+ P C+G+ + + C C +G+
Sbjct 30 CDPOYGFCEADNVCRCHVWEGFLC-----DRCVTP-----GCNVGCKEPMOICKDGN 80

Query 174 AGRECTINLDDCASRPCORGARCRD-RVHDFCLCPSGYGGKTCCLWL-PVPDPFTTVDTP 232
G-PC I+ C S PC C D ++C C G+ GK C+ B ++
Sbjct 81 DKGPCETIVRACTSTFCANNGTCVDEKGVQVCSCTPGFSKDCCKAGAPGVING 136

Query 233 LGFTSAVVVPATGPAPHSAGAGLLRISVKEVVRROEAGLGEPSLVALVFGALTAALVLA 292
V G A H+ S + F +T A
Sbjct 137 PQCHGGACVDDGGQASHA-----SCLCPGFSGMFCRIVA 172

Query 293 TVLLTLRMRGVCPGPC 311
T C P PC
Sbjct 173 T-----NSCTPFC 181

Score = 14.2 bits (25), Expect = 6.8, Method: Compositional matrix adjust.
Identities = 6/14 (42%), Positives = 7/14 (50%), Gaps = 0/14 (0%)

Query 317 HYAPACQDQECQVS 330
+ AC D E Q S
Sbjct 139 QHGGACVDDGGQAS 152

Select All Get selected sequences Distance tree of results

Cleavage of Membrane-Associated pref-1 Generates a Soluble Inhibitor of Adipocyte Differentiation

CYNTHIA M. SMAS, LI CHEN, AND HEI SOOK SUL*

Department of Nutritional Sciences, University of California, Berkeley, California 94720

Received 30 July 1996/Returned for modification 20 September 1996/Accepted 12 November 1996

pref-1 is an epidermal growth factor-like repeat protein present on the surface of preadipocytes that functions in the maintenance of the preadipose state. pref-1 expression is completely abolished during 3T3-L1 adipocyte differentiation. Bypassing this downregulation by constitutive expression of full-length transmembrane pref-1 in preadipocytes drastically inhibits differentiation. For the first time, we show processing of cell-associated pref-1 to generate both a soluble pref-1 protein of approximately 50 kDa that corresponds to the ectodomain and also smaller products of 24 to 25 kDa and 31 kDa. Furthermore, while all four of the alternately spliced forms of pref-1 produce cell-associated protein, only the two largest of the four alternately spliced isoforms undergo cleavage in the juxtamembrane region to release the soluble 50-kDa ectodomain. We demonstrate that addition of *Escherichia coli*-expressed pref-1 ectodomain to 3T3-L1 preadipocytes blocks differentiation, thus overriding the adipogenic actions of dexamethasone and methylisobutylxanthine. The inhibitory effects of the pref-1 ectodomain are blocked by preincubation of the protein with pref-1 antibody. That the ectodomain alone is sufficient for inhibition demonstrates that transmembrane pref-1 can be processed to generate an inhibitory soluble form, thereby greatly extending its range of action. Furthermore, we present evidence that alternate splicing is the mechanism that governs the production of transmembrane versus soluble pref-1, thereby determining the mode of action, juxtacrine or paracrine, of the pref-1 protein.

Adipose tissue is central to the maintenance of energy balance, and caloric intake in excess of energy utilization leads to obesity. Obesity may arise from increased size of individual adipose cells due to lipid accumulation or increased number of adipocytes arising from differentiation of adipose precursor cells to mature adipocytes. Overfeeding studies in rodents indicate a lifelong ability to make new fat cells in response to a high-fat diet and reveal that preadipocytes continue to undergo differentiation under the appropriate nutritional and hormonal cues throughout adulthood (10, 25, 26). Recent work reveals that the *ob* gene product, leptin, is an adipocyte-produced hormone involved in the regulation of appetite (17, 36, 51). Preadipocyte cell lines, such as 3T3-L1, differentiate *in vitro* in a process that biochemically and morphologically resembles *in vivo* adipocyte differentiation (14, 15). This spontaneous differentiation is accelerated by treatment of confluent preadipocytes with dexamethasone and methylisobutylxanthine (39). During differentiation, the fibroblastic preadipocyte becomes spherical and accumulates lipid; this is accompanied by dramatic alterations in the synthesis of cytoskeletal, extracellular matrix (ECM) proteins and those required for lipid metabolism, nutrient transport, and hormone responsiveness (reviewed in reference 42). Studies of genes expressed during adipocyte that differentiation, such as fatty acid binding protein (aP2), have demonstrated transactivation by C/EBP α (8, 19) and PPAR γ (47), whose ligand has recently been identified as 15-deoxy- $\Delta^{12,14}$ -prostaglandin J $_2$ (13, 24). Use of exogenous factors with positive or negative effects on adipogenesis have indicated that differentiation requires the appropriate combinatorial action of hormones, growth factors, and ECM (reviewed in reference 42). Preadipocytes, therefore, must integrate signals from the extracellular environment for differentiation to ensue. For example, while expression of C/EBP α and PPAR γ 2 promotes adipose-specific gene expres-

sion in cells not committed to the adipocyte lineage (20, 50), these effects generally require the addition of dexamethasone.

We originally identified preadipocyte factor 1 (pref-1) during a differential screening of a 3T3-L1 preadipocyte cDNA library designed to isolate genes that regulate adipogenesis (41). pref-1 is a transmembrane protein with epidermal growth factor (EGF)-like repeats in the extracellular domain, a juxtamembrane region, a single transmembrane domain, and a short cytoplasmic tail. The pref-1 transcript undergoes alternate splicing, with four major forms of the transcript detected in 3T3-L1 preadipocytes. The longest form, pref-1A, is most abundant; however, in-frame juxtamembrane deletions result in three additional transcripts. pref-1 is a unique regulatory molecule; it is expressed in preadipocytes, in contrast to the transcription factors, C/EBP and PPAR γ , that are only detected in conditions permissive for differentiation. pref-1 is readily detected in preadipocytes but is totally absent in mature fat cells, indicating complete downregulation of pref-1 during adipocyte differentiation. pref-1 mRNA levels are lower in 3T3-L1 preadipocytes than in the closely related but differentiation-defective 3T3-C2 cells. The level of pref-1 mRNA is decreased by treatment with a combination of the adipogenic inducing agents dexamethasone and methylisobutylxanthine (44) and by fetal calf serum, a component usually required for differentiation. Moreover, constitutive expression of pref-1 in 3T3-L1 preadipocytes inhibits their conversion to adipocytes as determined by cell morphology, level of adipocyte-expressed mRNAs, and degree of lipid accumulation. These observations indicate that pref-1 could maintain the preadipose phenotype and that pref-1 downregulation is integral to adipocyte differentiation.

The most striking feature of pref-1 is the presence of six tandem EGF-like repeats in the extracellular domain. The EGF-like repeat, first identified in epidermal growth factor, is a 35- to 40-amino-acid domain with conserved spacing of six cysteine residues. EGF-like domains are present in a number of molecules where they mediate protein-protein interaction to

* Corresponding author.

control cell growth and differentiation (1). A single EGF-like domain is the functional unit of EGF, transforming growth factor α (TGF α), and other growth factors that interact with the EGF receptor (32). Cleavage of a precursor transmembrane form at the amino and carboxyl termini of the EGF-like unit(s) releases the soluble growth factor (5). The extent of processing varies with the site of synthesis and is not requisite for biological activity (3, 33, 35, 46, 49). The importance of EGF-like domains in development is clearly illustrated by the *Drosophila* cell-fate determination proteins Notch (48) and Delta (28), transmembrane proteins that contain 36 and 9 EGF-like repeats, respectively. In contrast to the EGF-like growth factors, Notch and Delta function as transmembrane proteins; no processing of the ectodomain and/or release of EGF-like repeats occurs. Notch is a multifunctional receptor with pleiotropic effects. For example, interaction of the EGF-like repeats of Delta with those of Notch on adjacent cells transduces a lateral inhibitory signal during development of the neurogenic ectoderm (11). These studies suggest that EGF-like repeats function either as soluble ligands or as juxta-membrane-bound signalling or transducing molecules. Notably, the tandem arrangement of the pre-1 EGF-like repeats, as well as the interruption of its EGF-like repeats by introns (40), indicate that overall, pre-1 structure resembles that of *Drosophila* delta. pre-1 has been independently cloned as delta-like protein dkl (29) on the basis of its expression in several types of tumors. This observation, together with the function of pre-1 in adipocyte differentiation, has led us to hypothesize that in addition to the specific role of pre-1 in adipocyte differentiation, pre-1 could have a general role in maintaining the undifferentiated state. Indeed, pre-1 mRNA is detected in several embryonic tissues but not in their adult counterparts (41). The downregulation of pre-1 expression in differentiation, the inhibitory effects of its forced expression, and its EGF-like structural motif all predict that pre-1 may function in a manner analogous to that of Notch and Delta by interacting with EGF-like repeat proteins on adjacent cells or in the ECM to actively maintain the preadipose state.

In this report, we address whether transmembrane pre-1 undergoes processing to release a soluble factor and if the pre-1 ectodomain alone can generate the adipogenic inhibitory signal. By individual transfection of alternately spliced pre-1 cDNAs we show that membrane-associated pre-1 undergoes processing to produce several small soluble forms and a soluble product of 50 kDa that corresponds to the complete ectodomain. We find that two of the four alternately spliced pre-1 cDNAs do not generate this largest soluble form. When the pre-1 ectodomain produced as pre-1/glutathione S-transferase (GST) fusion protein is added to 3T3-L1 cells, their differentiation is drastically inhibited. Therefore, pre-1 not only functions as a transmembrane protein to affect adjacent cells but can act as a soluble inhibitor of adipocyte differentiation, with its mode of action, juxta-membrane or paracrine, determined by alternate splicing.

MATERIALS AND METHODS

Cell culture and transfection. 3T3-L1 cells and COS cells were maintained in Dulbecco's minimal essential medium (DMEM) with 10% fetal calf serum. For transfection, either COS-7 or COS-CMT cells were utilized as noted and seeded at 10^5 cells per 100-mm-diameter dish the day prior to transfection. Two micrograms of supercoiled DNA was transfected per dish utilizing DEAE-dextran (Stratagene). The pre-1 expression constructs utilized encompassed the open reading frame of pre-1 subcloned into either pcDNA1 or pcDNA1AMP (Invitrogen). For transfection of COS-7 cells, cells and DNA were kept in contact for 45 min, rinsed with phosphate-buffered saline (PBS), and incubated for 4 h in DMEM containing 10% fetal calf serum and 100 μ M chloroquine; the me-

dium was then changed to DMEM with 10% fetal calf serum. Unless otherwise stated, cells and medium were harvested at 72 h posttransfection. Transfection of COS-CMT cells was performed as described above except that cells were maintained in DMEM with 10% serum plus (RH Biosciences, Lenexa, Kans.) for 24 h following the onset of transfection and growth medium was supplemented with 100 μ M ZnCl₂ beginning at 24 h posttransfection.

Western blot analysis. Cell monolayers were rinsed twice with PBS and scraped into PBS containing 2 mM phenylmethylsulfonyl fluoride (PMSF). The cell suspension was subjected to three freeze-thaw cycles, and the crude membrane fraction was recovered by centrifugation at $13,000 \times g$ for 15 min at 4°C. The pellet was dissolved in lysis buffer (20 mM Tris-HCl [pH 7.4], 150 mM NaCl, 0.5% sodium deoxycholate, 1% Nonidet P-40 [NP-40], 1 mM EDTA, 1 mM PMSF) on ice for 30 min and clarified by brief spinning in a microcentrifuge, and the protein content was determined (Bio-Rad). The indicated amount of protein was loaded per lane in a sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gel and electrophoreted onto Immobilon polyvinylidene difluoride membranes (Millipore) with 10 mM 3-(carboxymethyl)-propylamine, acid-10% methanol transfer buffer. The pre-1 antibody was raised against a pre-1/TrpE fusion protein (41). For immunodetection of proteins, membranes were blocked for 1 h at room temperature in 5% nonfat dry milk-0.5% Tween 20 in PBS. Subsequent incubations and washes were conducted with 1% NET (145 mM NaCl, 5 mM EDTA, 0.25% gelatin, 0.05% Triton X-100, and 50 mM Tris-HCl [pH 7.4]). Detection of the antigen-antibody complex was accomplished via goat anti-rabbit immunoglobulin G-horse-radish peroxidase (HRP) conjugate (Bio-Rad), and signals were visualized by enhanced chemiluminescence (ECL) (Amersham) per manufacturer's instructions.

Metabolic labelling and immunoprecipitation. Seventy-two to ninety hours posttransfection, cell monolayers were rinsed with PBS and incubated in methionine and cysteine-free DMEM with 10% dialyzed fetal calf serum for 20 min. Following this, 200 μ Ci of ³⁵S TransExpress labelling mix (NEN) per ml was added. After the indicated labelling periods, medium was collected and monolayers were rinsed with PBS. Cells were either harvested or rinsed with DMEM with 10% fetal calf serum for the chase period. Following the chase period, medium was collected by sequential centrifugation at 1,100 and 17,000 $\times g$. For use in immunoprecipitation cell monolayers were harvested in 1% immunoprecipitation (IP) buffer (20 mM Tris-HCl [pH 7.4], 150 mM NaCl, 0.5% sodium deoxycholate, 1% NP-40, 1 mM EDTA, and 2 mM PMSF), and medium samples were adjusted to 1% IP buffer.

For immunoprecipitation, equal amounts of trichloroacetic acid-precipitable counts of ³⁵S-labelled cell lysates were brought to a volume of 125 μ l in lysis buffer and incubated with 10 μ l of antisera for 2 h on ice. Immune complexes were collected by incubation at 4°C with either fixed *Staphylococcus aureus* (Pansorbin; Calbiochem) for 15 min or Protein A-Sepharose (Pharmacia) for 1 h. Pellets were washed three times in radioimmunoprecipitation assay buffer (1% sodium deoxycholate, 1% NP-40, 0.1% SDS, 10 mM HEPES [pH 7.4], and 0.15 M NaCl). Samples were boiled in the presence of 2% (vol/vol) β -mercaptoethanol and fractionated on SDS-PAGE gels. For ³⁵S-labelled samples, gels were subjected to fluorography (Intensify; NEN) and exposed to Fuji RX film. For ³²P-labelled samples, gels were exposed to Fuji RX film with an intensifying screen.

In vitro transcription and translation. Full-length pre-1 cDNA in the EcoRI/XhoI site of the plasmid pcDNA1 was linearized at the unique 3' XhoI site. Capped, full-length pre-1 sense RNA was synthesized utilizing T3 polymerase (Stratagene). One microgram of synthesized transcript was used for in vitro translation with the incorporation of ³⁵S cysteine (NEN). Products were analyzed by SDS-10% PAGE and fluorography of dried gels (Intensify).

Posttranslational modification of pre-1 protein. For unimycin treatment of COS cells, cells were incubated at 72 h posttransfection with 10 μ g of tunicamycin or vehicle control during a 6-h metabolic labelling period. For enzymatic removal of N-linked carbohydrate and neuraminic acid, crude membrane fraction pellets of 3T3-L1 cells were dissolved in digest buffer (100 mM phosphate buffer [pH 7.0], 1% NP-40, and 200 μ M PMSF). Prior to digestion, 10 μ g of protein was denatured by boiling for 5 min in a final concentration of 0.5% SDS and 0.1 M β -mercaptoethanol in a total volume of 50 μ l. An additional 50 μ l of digest buffer was added per sample, samples were adjusted to 1 mM CaCl₂ for neuraminidase digestion and 2 U of N-glycanase (peptide N-glycosidase Fc Boehringer Mannheim) and 10 U of *Vibrio cholerae* neuraminidase (Boehringer Mannheim) were added as indicated for 3.5 h. Samples were analyzed by SDS-10% PAGE, and pre-1 protein was visualized by Western analysis.

Construction of c-Myc epitope-tagged pre-1 constructs. In tag the C-terminus of the pre-1 protein with the human c-Myc epitope, two oligonucleotides (coding strand, 5' GATCAGGACAGACCTGATCTCCGAGGAGGACCTCTA ATG 3'; noncoding strand, 5' GATCATCTATAGAGGTCTCTCTCGAGATCA CTTCTCTGCTC 3') were designed to encode the 10-amino-acid human c-Myc epitope recognized by the monoclonal antibody 9E10 (9) followed by an in-frame stop codon. The oligonucleotides were annealed by heating for 10 min at 79°C in 25 mM Tris (pH 7.6)-5 mM MgCl₂-25 mM NaCl; this was followed by a slow cooling to room temperature. The Myc tag was ligated into the pre-1/pcDNA1AMP expression constructs pre-1A and pre-1B at the C terminus of the pre-1 protein and the resulting fragments were ligated into the pre-1/pcDNA1AMP expression constructs.

Construction of P-tagged pre-1 and in vitro phosphorylation. A consensus phosphorylation site for the catalytic subunit of cAMP-dependent protein kinase,

encoding amino acids RRASV (termed herein P-tag), was inserted into the *NcoI* site that occurs at nucleotide 370 in the pref-1 cDNA sequence. This site was chosen because it occurs in EGF-like repeat two between the third and fourth cysteines, an area where spacing between cysteine residues is quite variable. Two oligonucleotides representing the coding (5'-CATGGCGCTCGCGGCTCTG TTG 3') and noncoding (5'-GACGACACGACACGCGGACGCG 3') strands with *NcoI*-compatible ends were annealed as described for the c-Myc oligonucleotides. The double-stranded product (P-tag) was ligated into the various c-Myc-tagged pref-1 expression constructs at the *NcoI* site.

Seventy-two hours posttransfection of COS-CMT cells, medium was collected by sequential centrifugation at 1,100 and 17,000 \times g, and the supernatant was acetone precipitated. Protein pellets corresponding to 2 ml of medium were collected by centrifugation, dried, and resuspended in bovine heart kinase phosphorylation buffer (20 mM Tris-HCl [pH 7.5], 100 mM NaCl, and 12 mM $MgCl_2$). A total of 125 μ l of this was used in the phosphorylation reaction that included 5 μ l of γ -ATP (3,000 Ci/mmol) and 50 U of heart muscle kinase (Sigma Chemical) in a final volume of 150 μ l. Following incubation at 4°C for 30 min, 850 μ l of stop solution (10 mM sodium phosphate [pH 8.0], 10 mM sodium pyrophosphate, 10 mM EDTA, and 1-mg/ml bovine serum albumin) was added, along with 110 μ l of 10 \times IP buffer (0.2 M Tris-HCl [pH 7.4], 1.5 M NaCl, 5% sodium deoxycholate, 10% NP-40, 10 mM EDTA, and 20 mM PMSE), and samples were divided equally for immunoprecipitation with 10 μ l of the indicated antisera.

pref-1/GST fusion protein production and 3T3-L1 cell differentiation. EcoRI/BamHI-digested pGEX2TK (Pharmacia Biotech) and PCR-amplified fragments of pref-1 were used to generate expression vectors for pref-1/GST fusion proteins. GST and pref-1/GST, corresponding to the full pref-1 extracellular domain minus the signal sequence (amino acids 8 through 299), were expressed in BL-21 *Escherichia coli* and purified by affinity binding to glutathione agarose beads (Pharmacia Biotech). The proteins eluted by 5 mM reduced glutathione were dialyzed against 1 \times PBS, mixed with DMEM containing 0.5% FBS, and filter sterilized through a Millex 10- μ m-pore-size filter (Millipore). At confluence, 3T3-L1 preadipocytes were treated for 48 h with 1 μ M dexamethasone and 0.5 mM methylxanthine (dex/mx). Control GST or pref-1/GST proteins were added at the start of the differentiation protocol at a concentration of 50 nM. This concentration was maintained by the addition of proteins at subsequent medium changes. The concentration of purified proteins was determined by multiplying the purity of the protein determined by Coomassie blue staining of SDS-PAGE gels by the total protein concentration. Antisera were inoculated with equal volumes of the fusion protein before being added to the medium at a final dilution of 1:100. The effect of pref-1 inhibition could first be observed 2 days after the start of the differentiation protocol. At 5 days postinitiation of differentiation, cells were stained for lipid with Oil Red O and photographed, and RNA was extracted from parallel cultures and subjected to Northern analysis as previously described (41) utilizing 32 P-labeled cDNA for fatty acid synthetase, C/EBP α , steroloyl coenzyme A desaturase, and fatty acid binding protein and a labeled PPAR- γ cDNA that detects both the PPAR- γ 1 and PPAR- γ 2 transcripts.

RESULTS

Cleavage yields a residual C-terminal 25-kDa cell-associated pref-1. To begin characterizing the pref-1 protein, we generated antibodies against an *E. coli*-expressed TrpE/pref-1 fusion protein. In 3T3-L1 cell lysates, a minimum of seven discrete protein bands of approximately 45 to 55 kDa are detected by the pref-1 antibody. These are abolished by preincubation of the pref-1 antisera with TrpE/pref-1 fusion protein but not with TrpE protein alone (43). Given the complex pattern of pref-1 protein in preadipocytes, partly due to expression of at least four alternate transcripts, detailed analysis of specific pref-1 protein isoforms is inherently difficult. To overcome these limitations we expressed pref-1 in COS cells, a cell type that lacks endogenous pref-1 and that has been extensively utilized to address protein structure and function. When we transfected full-length pref-1 we observed that, in addition to full-length transmembrane pref-1, a 25-kDa cell-associated protein was specifically detected by pref-1 antibody. Since pref-1 was previously known to exist only in transmembrane form, the appearance of this 25-kDa membrane-associated protein was the first indication we had that pref-1, in addition to its transmembrane location, might exist in a soluble form; the 25-kDa protein could correspond to residual pref-1 after cleavage and release of some region of the ectodomain and would thus be predicted to contain the pref-1 cytoplasmic domain.

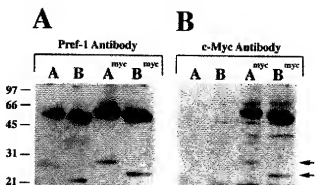


FIG. 1. Western analysis of c-Myc-tagged pref-1. Twenty-five micrograms of protein from COS-7 cells expressing pref-1A (lanes A), pref-1B (lanes B), or forms tagged with the c-Myc epitope (lanes A^{myc} and B^{myc}) were fractionated on SDS-10% PAGE. (A) Western analysis utilizing a 1:15,000 dilution of pref-1 primary antibody and a 1:2,000 dilution of goat anti-rabbit secondary antibody followed by ECL detection. (B) The same blot shown in panel A stripped and reprobed with a 1:20,000 dilution of the 9E10 primary antibody and a 1:2,000 dilution of goat anti-mouse secondary antibody followed by ECL detection. Molecular mass markers in kilodaltons are on the left, and the arrows at the right indicate the positions of the 25-kDa pref-1A and 21-kDa pref-1B protein bands.

To determine if the 25-kDa cell-associated protein corresponds to the C terminal cytoplasmic domain, we added a 10-amino-acid human c-Myc epitope tag to the extreme C terminus of cDNA expression constructs for the two longest forms of pref-1, pref-1A and pref-1B. Myc-tagged and unmodified versions of pref-1A and pref-1B were transfected into COS cells, and crude membrane fraction proteins were analyzed by Western blotting. The pref-1 antibody detects the full-length 55-kDa pref-1A and the full-length 51-kDa pref-1B in the membrane fraction. In addition, a 25-kDa protein results upon pref-1A expression, and a 21-kDa protein results by pref-1B expression (Fig. 1A). This 4-kDa size difference of the pref-1A and pref-1B proteins reflects the membrane-proximal 153-base deletion in pref-1B arising by alternate splicing. The Myc-tagged versions of pref-1A and pref-1B are also recognized. In each case the addition of the Myc tag increases the molecular mass of the pref-1 bands by 1 kDa, i.e., the size of tag. This is most apparent for the 25-kDa pref-1A and the 21-kDa pref-1B proteins. Their size increases upon addition of the Myc tag indicate that they contain the C terminus of the pref-1 protein. Reprobing of the same membrane with the 9E10 antibody specific for the Myc epitope (9) shows that only the Myc-tagged, and not the native forms, of pref-1A and pref-1B are specifically recognized (Fig. 1B). Furthermore, the identical 25-kDa pref-1A and 21-kDa pref-1B proteins are recognized by the pref-1 and the 9E10 antibodies (Fig. 1B). Given the 1-kDa size increase that occurs with the addition of the Myc tag, and the recognition of these bands by both antibodies, we conclude that full-length membrane pref-1 is probably cleaved to a residual membrane-associated protein of 25 kDa for pref-1A and 21 kDa for pref-1B and that this protein contains the pref-1 cytoplasmic domain.

The pref-1 ectodomain is cleaved to a soluble factor. Our Myc tag studies indicate that membrane pref-1 undergoes cleavage. To detect the pref-1 cleavage product in the medium and address pref-1 processing in more detail we expressed pref-1A in COS cells and performed pulse-chase analyses. At the end of the 30-min pulse period, a pref-1 protein of 55 kDa is detected by immunoprecipitation of the cell lysate with pref-1 antibody (Fig. 2A). By 7 h postsynthesis, the majority of membrane-associated pref-1A has been turned over, and it is

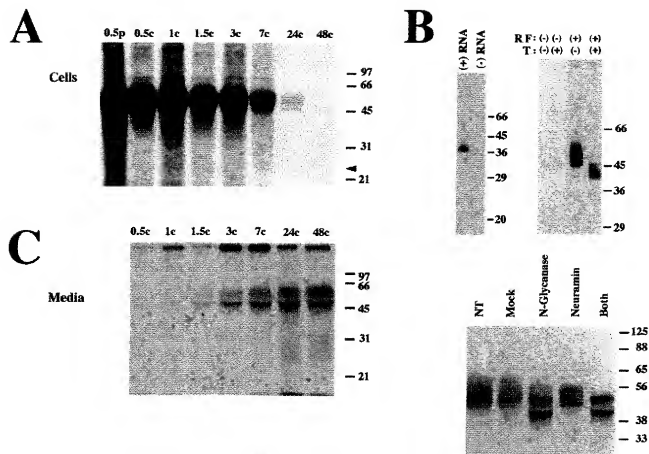


FIG. 2. Analysis of pref-1 processing. (A) Pulse-chase analysis of cellular pref-1. pref-1A-expressing COS-CMT cells were pulse-labelled with [35 S]cysteine and methionine for 30 min (0.5p) and subjected to the indicated chase periods (c) in hours followed by immunoprecipitation of cell lysates with pref-1 antibody and SDS-PAGE. Cells were harvested at indicated time points. Normal sera controls indicate that the approximately 58-kDa doublet band seen at 24 and 48 h is nonspecific (43). The exposure time for cell-associated pref-1 is approximately one-fifth that for soluble pref-1 shown in panel C. The arrowhead indicates the position of the 25-kDa product. (B) Posttranslational modification of pref-1. The left part of the panel shows results of *in vitro* translation of *in vitro*-transcribed pref-1 RNA (+), or a no RNA control (-), in the presence of [35 S]cysteine. The right part of the panel shows results from COS cells transfected with the correct (RF+) or reverse (RF-) orientation of the pref-1 open reading frame and which were subjected to metabolic labelling with [35 S]cysteine and methionine in the presence (T+) or absence (T-) of tunicamycin, immunoprecipitated with pref-1 antibody, and analyzed by SDS-PAGE. The lower part of the panel shows results from denatured crude membrane fraction protein from 3T3-L1 cells which was either not treated (NT), incubated without addition of enzyme (mock), or treated with N-glycanase, neuraminidase, or both N-glycanase and neuraminidase for 3.5 h. Following digestion, 50- μ g samples were fractionated on SDS-10% PAGE; gels and subjected to Western analysis using pref-1 antisera at a dilution of 1:800 and a 1:2,000 dilution of goat anti-rabbit HRP secondary antibody. (C) Pulse-chase analysis of soluble pref-1. pref-1A-expressing COS-CMT cells were pulse-labelled with [35 S]cysteine and methionine for 30 min and subjected to the indicated chase periods (c) in hours followed by immunoprecipitation of medium with pref-1 antibody and SDS-PAGE. Samples of medium were collected at indicated time points. The exposure time for the samples was approximately five times longer than that for the pulse-chase analysis of cell-associated pref-1 shown in panel A. Molecular mass markers in kilodaltons are shown on the right.

undetectable at 48 h. At the same time a 25-kDa product (Fig. 2A), which is the size of the residual membrane and cytoplasmic domain of pref-1 detected by Western blotting shown in Fig. 1, is in the cell lysate. However, the prominence and relative ratio of the 25-kDa form to that of full-length membrane-associated pref-1 differs in our Western blot versus pulse-chase analyses. This apparent discrepancy may be because whereas Western analysis likely reflects steady-state actual molar ratios, the signals of the metabolically labelled proteins are based on their content of methionine and cysteine and reflect a single time point of synthesis. Since the majority of the pref-1 extracellular domain consists of six tandem EGF-like repeats, each containing six cysteine residues, the intensity of bands detected by pulse-chase studies would be skewed toward full-length pref-1 in contrast to the residual membrane-associated 25-kDa form.

In these analyses, the cell-associated 55-kDa pref-1 appears as a broader signal than that shown in Fig. 1. Although this could be attributable to the detection technique used, immunoprecipitation versus Western analysis and gel resolution, the diffuse nature of the 55-kDa cell-associated pref-1 shown in Fig. 2A suggests posttranslational modification of the protein. There are three consensus sites for N-linked glycosylation in the pref-1 extracellular domain. We determined the size of the pref-1 primary translation product and utilized tunicamycin, an inhibitor of N-linked glycosylation to assess whether pref-1 protein in transfected COS cells contains N-glycan (Fig. 2B). *In vitro* translation results in an approximately 39-kDa protein, which is in agreement with the predicted size of the pref-1 primary translation product (Fig. 2B, left panel). Metabolic labelling and immunoprecipitation of pref-1-expressing COS cells reveals that cell-associated pref-1 protein is reduced to a

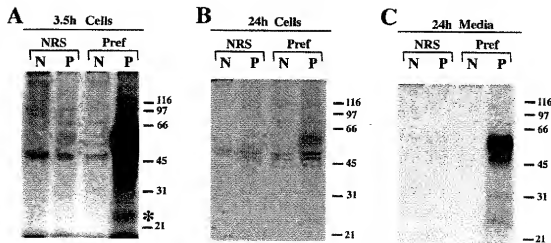


FIG. 3. Detection of soluble forms of pref-1 in conditioned medium. Nontransfected (N) or COS-CMT cells transfected with pref-1A (P) were metabolically labelled with 35 S for 3.5 h, cells and medium were subjected to immunoprecipitation with normal rabbit sera (NRS) or pref-1 antisera (Pref), and products were fractionated on SDS-10% PAGE gels. (A) Cells harvested following the 3.5-h labelling period. The asterisk indicates a band that may correspond to a residual 25-kDa pref-1 protein associated with the cytoplasmic membrane (Fig. 1 and 2A). (B) Cells harvested 24 h after the onset of the 3.5-h labelling period. (C) Medium harvested 24 h after the onset of the 3.5-h labelling period. Molecular mass markers in kilodaltons are on the right.

more discrete band of 45 kDa in the presence of tunicamycin. These bands are not present when an expression construct containing the opposite orientation of the pref-1 reading frame is employed (Fig. 2B, right panel). This indicates that all of the heterogeneous cell-associated proteins we detect correspond to various forms of pref-1. To further address this, crude membrane preparations of 3T3-L1 preadipocytes were treated with *N*-glycanase and neuraminidase, followed by Western analysis. No treatment and mock treatment show multiple discrete pref-1 protein bands. Digestion with *N*-glycanase, neuraminidase, or a combination confirms pref-1 is a glycoprotein that contains *N*-linked oligosaccharide and sialic acid (Fig. 2B, lower panel). The presence of sialic acid in pref-1 may therefore explain the 6-kDa size difference between the *in vitro*-translated product and pref-1 protein present in tunicamycin-treated cells. We conclude that the heterogeneous nature of pref-1 protein is due to posttranslational modifications that occur within 30 min of synthesis.

Pulse-chase analysis of the medium (Fig. 2C) demonstrates that a soluble 50-kDa form of pref-1 appears 1.5 h postsynthesis and accumulates thereafter. In addition, a diffuse signal between 21 and 31 kDa is present in the medium at 24 h. The increase in soluble pref-1 in the medium with a concomitant decrease in the membrane-associated form (Fig. 2A) is consistent with a precursor-product relationship and indicates cell-associated pref-1 is processed to release soluble products. This does not necessarily indicate that all of membrane-associated pref-1 undergoes processing; the decrease in the 55-kDa cell-associated form over time is likely due to the combined effects of cleavage to soluble forms and recycling and/or turnover of membrane pref-1. To further address the nature of these smaller proteins in the medium, a longer labelling period was used. COS cells were transfected with pref-1A, and cells were harvested 3.5 h after labelling or cells and medium were harvested 24 h after the onset of labelling. After the 3.5-h labelling period a pref-1 band of approximately 55 kDa, corresponding to full-length pref-1A, is detected in pref-1-transfected cells (Fig. 3A). It is not present in nontransfected controls nor is it detected with normal rabbit sera. An additional band (Fig. 3A) may correspond to the residual cytoplasmic membrane-associated

25-kDa pref-1 protein noted in Fig. 1 and 2A. As in the pulse-chase analysis, cellular pref-1 protein is barely detectable 24 h postlabelling (Fig. 3B). However, this longer labelling identified, in addition to the prominent 50-kDa soluble form, 24- to 25-kDa and 31-kDa proteins in the medium (Fig. 3C). The diffuse nature of the 24- to 25-kDa doublet suggests it could arise by differential posttranslational modification of the same polypeptide backbone. The low amounts of these smaller soluble forms may indicate a slow cleavage event due to a limiting proteolysis system. As the immunoprecipitation analysis revealed the cell-associated and soluble pref-1 to be close in size, 55 kDa versus 50 kDa, and since these analyses were performed on separate SDS-PAGE gels, to confirm this size difference we directly compared the size of cell-associated and soluble pref-1 by resolving them in adjacent lanes of an SDS-PAGE gel. Figure 4 shows the result of this size comparison analyzed by Western blotting pref-1 protein is not detected in nontransfected COS cells whereas transfection of pref-1A results in an approximately 55-kDa cell-associated pref-1 protein and an approximately 50-kDa form in the medium. These findings are in agreement with the metabolic labelling results shown in Fig. 2 and 3 and taken together indicate that full-length cell-associated pref-1 can undergo processing to release a 50-kDa soluble form.

Localization of cleavage and regulation of soluble pref-1 production by alternate splicing. Identification of multiple soluble forms of pref-1 and a 25-kDa cell-associated form indicates that membrane pref-1 is subject to two cleavage events. Based on the 50-kDa molecular mass for the large soluble form, this cleavage event would occur near the cell membrane. The cleavage event that generates smaller soluble pref-1 would be predicted to occur at a more membrane-distal site. To study the generation of the soluble pref-1 in more detail, we used two approaches: (i) addition of a phosphorylation site tag (P-tag) to the pref-1 extracellular domain and (ii) determination of the effect of various juxtamembrane deletions on the appearance of soluble pref-1. We hypothesized that processing from the N terminus may generate the smaller soluble forms of pref-1 detected by metabolic labelling. To determine which portion of the extracellular domain of pref-1 is released as the soluble

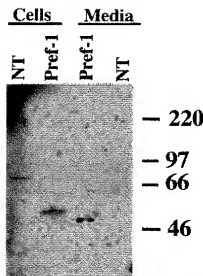


FIG. 4. Western analysis of cell-associated and soluble pref-1. Cells and conditioned medium were harvested from COS-CMT cells transfected with pref-1A (Pref-1) or nontransfected (NT) controls. Fifteen micrograms of the cell lysate and 5 μ l of conditioned medium were fractionated on SDS-10% PAGE gels, subjected to Western analysis using a 1:15,000 dilution of pref-1 primary antibody and a 1:5,000 dilution of goat anti-rabbit-HRP secondary antibody, and products were visualized by ECL.

form(s), a consensus phosphorylation site (P-tag) for cAMP-dependent protein kinase was added near the N terminus of the pref-1 extracellular domain. To minimize the effects of the addition of six amino acids on overall structure, the P-tag was inserted in the second EGF-like repeat between the third and fourth cysteines, an area with variable cysteine spacing. P-tagged pref-1A was expressed in COS cells and the medium was in vitro phosphorylated and immunoprecipitated (Fig. 5). We detected a phosphorylated protein of 50 kDa, the same soluble product noted by metabolic labelling; given its size this protein likely corresponds to the full ectodomain. The doublet of 24 to 25 kDa is also observed by use of the P-tag. These proteins therefore contain the second EGF repeat and thus probably the N-terminal region of pref-1. They are not detected in nontransfected COS cells nor when normal sera or an unrelated antiserum was used in immunoprecipitation. We therefore predict that a pref-1 processing event occurs at a site C terminal to the P-tag to generate the N-terminal, P-tagged 24- to 25-kDa doublet. This membrane-distal event would also explain our detection of the 25-kDa residual cell-associated pref-1 which is apparent upon expression of pref-1A as shown in Fig. 1. The sizes of the soluble forms detected by metabolic labelling and P-tag are identical. The differences observed in the relative ratio of the 50-kDa to the 24- to 25-kDa soluble form may be attributed to inherent differences in the two detection methods. While metabolic labelling at cysteine residues, abundant in the EGF-like repeat motif, follows a population of pref-1 synthesized during a specific period at 72 h posttransfection, by in vitro phosphorylation each pref-1 molecule is labelled at a single P-tag site. The signals of the various soluble forms determined by in vitro phosphorylation likely reflect a steady-state level of their molar ratios accumulated from 24 to 72 h posttransfection.

The four alternatively spliced forms of pref-1 that we previously identified in 3T3-L1 preadipocytes, and which have various in-frame extracellular juxtamembrane deletions, provided

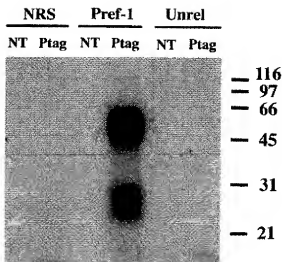


FIG. 5. Analysis of P-tagged pref-1 in medium. Conditioned medium collected from nontransfected (NT) COS-CMT cells or COS-CMT cells transfected with the P-tagged version of pref-1A (Ptag) was in vitro phosphorylated with γ -³²P and immunoprecipitated with either normal rabbit sera (NRS), pref-1 antisera (Pref-1), or antisera raised against an unrelated TrpE fusion protein (Unrel). Immunoprecipitates were fractionated by SDS-10% PAGE and subject to autoradiography. Molecular mass markers in kilodaltons are on the right.

a system in which to test our hypothesis that the 50-kDa soluble pref-1 derives from an extracellular membrane-proximal cleavage. The structures of these alternate forms are depicted in Fig. 6B. Transfection of each of the four major alternate forms of the pref-1 cDNA results in membrane-associated pref-1 proteins whose molecular masses decrease in correspondence to their respective deletions (43). To address soluble pref-1 production, the four P-tagged alternately spliced forms were expressed in COS cells. The medium was subject to in vitro phosphorylation at the P-tag site, immunoprecipitation, and SDS-PAGE analysis. Strikingly, whereas each isoform expresses the 24- to 25-kDa doublet in the medium, the large soluble form is produced only by pref-1A and pref-1B; little if any large soluble pref-1 is generated by the two alternately spliced isoforms with larger juxtamembrane deletions, pref-1C and pref-1D (Fig. 6A). These data reveal that the cleavage that generates large soluble pref-1 occurs within a sequence common to pref-1A and pref-1B. Furthermore, the observation that pref-1B results in the large soluble form and pref-1C does not indicates that the sequence present in pref-1B, but deleted in pref-1C, contains the membrane-proximal processing site for the generation of the large soluble pref-1. This localizes the cleavage event to within the 22-amino-acid juxtamembrane sequence PEOHILKYSMKELNKPSTPLTE (Fig. 6C). Interestingly, our localization of the membrane-proximal cleavage to within the sequence PEOHILKYSMKELNKPSTPLTE agrees with the protein sequence of fetal antigen 1 (FA1), reported during the course of our experiments. FA1 is a circulating fetal protein with undetermined function that likely corresponds to the complete extracellular domain of human pref-1 (23). The N terminus of FA1 begins after the pref-1 signal sequence, and although the extreme C terminus of FA1 has not been unambiguously assigned, it falls within the 22-amino-acid sequence that we determined contains the membrane-proximal cleavage site for the release of the 50-kDa soluble pref-1. Although no consensus processing sites are present in this 22-amino-acid sequence, the sequence is nota-

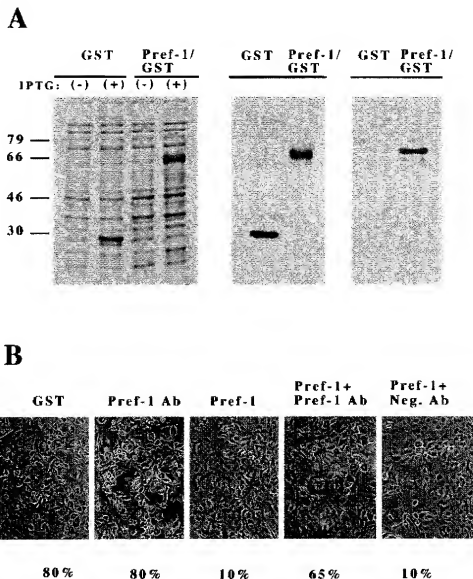


FIG. 8. Production and activity of pref-1/GST fusion protein. (A) Expression, purification, and Western analysis of pref-1/GST fusion protein. Left, Coomassie blue-stained SDS-PAGE gel of total protein from uninduced (–) and IPTG-induced (+) BL-21 *E. coli* harboring either the GST or pref-1/GST expression constructs; middle, affinity-purified 29-kDa GST and the 63-kDa pref-1/GST fusion protein on a Coomassie blue-stained SDS-PAGE gel; right, Western analysis of purified GST and pref-1/GST using pref-1 antibody. Molecular mass markers in kilodaltons are on the left. (B) The inhibitory effect of pref-1/GST on 3T3-L1 differentiation assessed by Oil Red O staining of cellular lipid. Additions to the standard dex/mix differentiation treatment are noted above photomicrographs. GST, GST protein; Pref-1 Ab, antibody directed against a pref-1/TrpE fusion protein; Pref-1, pref-1/GST fusion protein; Neg. Ab, antibody directed against an unrelated TrpE fusion protein. The percentages of lipid-containing cells are indicated below the photomicrographs.

fusion protein in *E. coli*. Cells harboring the GST or pref-1/GST expression construct show an identical pattern of proteins upon Coomassie blue staining of SDS-PAGE gels. Induction of protein expression with isopropyl- β -D-thiogalactopyranoside (IPTG) results in proteins of the size predicted for GST alone (29 kDa) or pref-1/GST (63 kDa); these are the most abundant proteins detected (Fig. 8A, left panel). Coomassie blue staining of soluble fusion proteins after affinity binding to glutathione agarose beads shows a single band, indicating purification to near homogeneity (Fig. 8A, middle panel). Western analysis reveals that the pref-1/GST fusion protein but not GST alone is specifically detected by pref-1 antibody (Fig. 8A, right panel). To test the effect of soluble pref-1 on adipocyte differen-

tiation, confluent 3T3-L1 preadipocytes were treated with dex/mix to initiate differentiation. The medium was supplemented with either the GST protein or the pref-1/GST fusion protein. Additionally, to address the specificity of the effects of pref-1/GST, pref-1 antibody or an antibody against an unrelated TrpE fusion protein was utilized. After 5 days, cells were fixed and stained with Oil Red O, and the degree of adipocyte differentiation was judged by cell morphology and the percentage of lipid-containing cells (Fig. 8B). Addition of either the GST protein or pref-1 antibody had no discernable effects; 80% of cells differentiated to adipocytes as indicated by high lipid content and rounded appearance. Addition of pref-1/GST fusion protein markedly inhibited differentiation, and these cells

TABLE 1. Concentration-dependent inhibition of 3T3-L1 differentiation by GST-pref-1 fusion protein*

Protein concentration (nM)	% Adipocyte conversion	
	Expt 1	Expt 2
0	>70	>70
5	>70	>70
10	20-50	30-60
25	<20	20-50
50	<20	<20
100	<20	<20

* Confluent 3T3-L1 cells in quadruplicate dishes were subject to dex/mix treatment in the presence of the indicated concentrations of fusion protein. Six days after induction cells were stained for lipids with Oil Red O and examined microscopically for percentage of adipocyte conversion. The average of four dishes is indicated.

had very little lipid accumulation and maintained fibroblast morphology with only 10% of the cells differentiating. Furthermore, the inhibitory effects of the fusion protein on adipocyte differentiation are attenuated by preincubation of pref-1/GST with antiserum against pref-1. These cultures show 65% differentiation, whereas preincubation with an unrelated control serum does not affect the inhibitory action of pref-1/GST as evidenced by only 10% differentiation. This indicates that the inhibitory effects of the pref-1/GST fusion protein are specifically due to the pref-1 ectodomain. To address whether there is a dose-response effect of pref-1 action, we tested the inhibitory action of pref-1 at protein concentrations of 0, 5, 10, 25,

50, and 100 nM. As is shown in Table 1, the inhibitory effects of pref-1 are first noted at 10 nM, and maximum inhibition is observed at 50 nM. These data, as well as the blocking effects of pref-1 antibody shown in Fig. 8B, are consistent with the existence of a specific pref-1 receptor. However, the biological nature of the assay system, namely the inhibition of adipocyte differentiation, limits more detailed determination of the kinetics of pref-1 interaction with its predicted receptor. Such analyses await the identification and isolation of the pref-1 receptor by interaction cloning or other methods.

We next addressed the inhibition of adipocyte differentiation in detail at the molecular level. The effects of soluble pref-1 on the level of adipocyte-expressed RNAs was determined and correlated with morphological evidence of adipocyte differentiation. Cells were treated with dex/mix alone or supplemented with GST or pref-1/GST and stained for lipid with Oil Red O 5 days after initiation of differentiation (Fig. 9A); we observed the same inhibitory effects for pref-1 which are shown in Fig. 8B. Northern analysis for five adipocyte-expressed mRNAs reveal that compared to cells differentiated with the standard dex/mix treatment or with the addition of GST protein, pref-1-treated cells have only 20% of the levels of the terminal marker mRNAs for fatty acid synthase, stearoyl coenzyme A desaturase, and fatty acid binding protein (Fig. 9B). Moreover, the levels of mRNA for C/EBP α and PPAR γ are similarly decreased, indicating the inability of cells to express these transcription factors in the presence of soluble pref-1. This suggests that the inhibitory effects of soluble pref-1 are exerted early in differentiation. The results indicate that

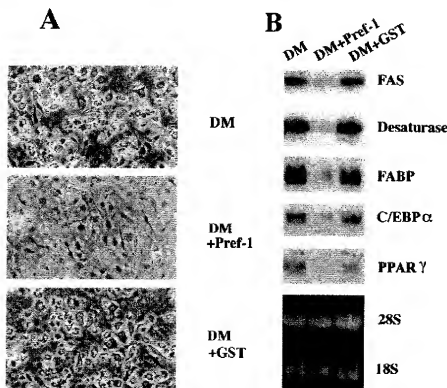


FIG. 9. Soluble pref-1 inhibits adipocyte differentiation. (A) Confluent 3T3-L1 preadipocytes were subject to standard *in vitro* differentiation conditions (DM) or supplemented with either the pref-1/GST fusion protein (DM+Pref-1) or GST control (DM+GST) throughout the course of differentiation. At 5 days after initiation of differentiation, cultures were stained with Oil Red O and photographed. Typical micrographs of the cultures are shown. (B) Ten micrograms of total RNA from parallel cultures were subject to Northern blot analysis using the indicated ³²P-labelled cDNA probes. The PPAR γ signal appears as a doublet since both the γ 1 and γ 2 isoforms are detected. Representative ethidium bromide staining of the Northern gel is shown at the bottom.

the pref-1 ectodomain alone, corresponding to the soluble form we detect in conditioned medium, is sufficient for the inhibitory effect of pref-1 in adipocyte differentiation. This further suggests that the pref-1 molecule, in either transmembrane or soluble form, probably functions as a ligand to initiate and/or maintain signals inhibitory to adipogenesis.

DISCUSSION

pref-1 exists in both transmembrane and soluble forms. We demonstrate by pulse-chase analyses and *in vitro* phosphorylation at the P-tag site that full-length pref-1 undergoes cleavage at a membrane-proximal site to release an N-terminal soluble product of 50 kDa. This soluble form inhibits adipocyte differentiation. The differential effects of alternate splicing on the production of the 50-kDa soluble pref-1 predicts cleavage occurs extracellularly near the transmembrane domain at a membrane-proximal site within the 22-amino-acid sequence PEOHILKYSMKELNKSPTLLTE. This agrees with the protein sequence of the human fetal protein FA1, reported during the course of our studies, that corresponds to the pref-1 extracellular domain. The simplest interpretation of our data is that the spliced-out sequence removes a processing site. This has strong similarities to the effect of alternate splicing in the c-kit ligand where the KL-1 form is processed while the alternately spliced KL-2 form is not efficiently cleaved due to a juxtamembrane deletion encompassing the preferred processing site (12, 21). This 22-amino-acid sequence does not contain any recognizable motifs such as the basic residues that are processing sites for kex2/furin proteases (18) or the small apolar amino acids where cleavage of TGF α occurs (31). It is of interest to note that the splicing event removes portions of the juxtamembrane region, including sequences reminiscent of a leucine zipper.

These findings place pref-1 into that class of proteins which can act either as transmembrane or soluble molecules. Among EGF-like repeat proteins, ectodomain processing and release has been demonstrated only for those growth factors that function through the EGF receptor and related receptors. Transfection studies with EGF and TGF α have allowed detailed analysis of their processing from transmembrane precursors; however, processing is not requisite for their biological activity. Membrane-anchored forms of EGF and TGF α bind and activate the EGF receptor (2, 32). The full 160-kDa pro-EGF produced by the kidney is active (37). Ectodomain release also occurs for transmembrane molecules other than the EGF-like repeat growth factors, including the c-kit ligand (7, 21), tumor necrosis factor 1 receptor (16, 27), and the β -amyloid precursor protein (4). While our data indicate that the 50-kDa soluble form of pref-1 corresponds to the full ectodomain as the result of membrane-proximal cleavage, formulation of a complete model for the generation of the minor soluble forms of pref-1 is not yet possible. In addition to a prominent soluble form of 50 kDa, we also find that the 24- to 25-kDa form contains the P-tag placed near the pref-1 N terminus. We can speculate that subsequent cleavage of the 50-kDa form of soluble pref-1 at the membrane-distal site to generate the 24- to 25-kDa soluble pref-1 could serve as a mechanism to inactivate the larger soluble form or otherwise modulate its activity. However, as we have not yet clearly delineated which portion of the pref-1 ectodomain the smaller soluble proteins derive from, we cannot at this time address with any certainty their function. Nevertheless, a membrane-distal processing event would be predicted to occur at a site C terminal to the P-tag inserted in the second EGF-like repeat. This would release the 24- to 25-kDa soluble form. We predict that this

membrane-distal event also generates the 25-kDa residual cell-associated pref-1 that we determined by Myc-epitope tagging to contain the pref-1 cytoplasmic domain. With the assumption that the size of this 25-kDa residual cell-associated pref-1 is attributable solely to primary amino acid sequence, the 25-kDa residual cell-associated protein would correspond to the extreme C terminus of pref-1 up to EGF-like repeat five. Inspection of the primary amino acid sequence of pref-1 within the region bordered by the P-tag and the transmembrane domain reveals an area of small apolar amino acids, Val-Ala-Ala, between the fourth and fifth EGF-like repeats. This is similar to the cleavage site(s) used for the release of mature soluble EGF, TGF α , and KL-1 from transmembrane precursors (30, 34). Preliminary studies indicate site-directed mutagenesis of the pref-1 Val-Ala-Ala sequence alters the amount and appearance of soluble pref-1 (43).

Functional implications of pref-1 processing. The work presented demonstrates that the pref-1 ectodomain/GST fusion protein, which corresponds to the 50-kDa soluble form, inhibits adipocyte differentiation, as we have previously shown for the membrane-associated form. Since the 50-kDa soluble form of pref-1 has inhibitory activity similar to that of the full-length membrane-associated form, release of the pref-1 ectodomain as a soluble factor allows switching between two active forms of pref-1, thereby regulating its range of action. Therefore, pref-1 not only functions in a juxtacrine manner as a transmembrane protein to affect adjacent cells but can have paracrine actions as a soluble inhibitor of adipocyte differentiation. We have confirmed the inhibitory effects of soluble pref-1 by treating confluent 3T3-L1 preadipocytes with dex/mix in the presence of conditioned medium from transfected COS cells. Following a 2-day dex/mix treatment, cells were maintained in 50% fresh growth medium-50% conditioned medium. While cells treated with conditioned medium from mock-transfected COS cells differentiated well, as judged by the number of lipid-containing cells, conditioned medium from pref-1A-transfected COS cells drastically reduced adipocyte differentiation (43). Thus, use of two different approaches, GST fusion protein and conditioned medium, demonstrates the inhibitory action of soluble pref-1 and indicates that the bioactivity of the GST fusion protein is similar to that produced by COS cells. Although both the transmembrane and soluble pref-1 are active in the inhibition of adipocyte differentiation, future studies may reveal finer distinctions in their respective functions, as demonstrated for the kit ligand where the soluble factor does not fully substitute for the actions of membrane-bound kit ligand *in vivo* (12). These inhibitory effects observed with pref-1A-conditioned medium are additional evidence for an *in vivo* role of soluble pref-1 in the regulation of adipocyte differentiation. Moreover, we have observed that treatment of 3T3-L1 preadipocytes with conditioned medium from COS cells transfected with pref-1A markedly inhibits adipocyte differentiation, while conditioned medium from COS cells transfected with the most deleted alternate form, pref-1D, does not affect adipocyte differentiation (43). We therefore hypothesize that the mode of function, juxtacrine or paracrine, depends on the alternate pref-1 transcript expressed.

The temporal expression of genes during adipocyte differentiation suggests a hierarchy of regulatory events. Based on expression pattern and transfection studies, C/EBP and PPAR γ have been shown to be central to adipogenesis. However, factors such as cell confluence/growth arrest, fetal calf serum, dexamethasone, and an ECM environment conducive to adipocyte differentiation may govern expression and action of these transcription factors. The absolute downregulation of pref-1 during adipocyte conversion and the inhibitory effects of

forced pref-1 expression in preadipocytes suggest it has a unique regulatory function in this process. In conditions under which preadipocytes normally differentiate, addition of soluble pref-1 prevents expression of both PPAR γ and C/EBP α , the regulatory molecules that transactivate adipocyte genes and lead to adipogenesis. This is consistent with the concept that downregulation of pref-1 is a prerequisite for C/EBP α and PPAR γ induction and adipocyte differentiation. Our experiments here suggest that, via the generation of a soluble inhibitory form, pref-1 is likely to have a wider range of function than was first predicted on the basis of its synthesis as a transmembrane protein. The inhibitory effects of fibronectin (45) and collagen (22) on adipocyte differentiation indicate cytoskeletal and/or ECM remodelling is requisite for adipocyte differentiation. By analogy, pref-1, as either a transmembrane or soluble protein, may exert its inhibitory effects through interaction of its EGF-like repeats with EGF-like repeats present in cell surface or ECM components to maintain the preadipose phenotype. It is intriguing given its structural similarities to the Notch-Delta family, that pref-1 is processed to generate soluble forms. Work presented here does not rule out the possibility that transmembrane pref-1 may act as a receptor to transduce inhibitory signals. However, the fact that the pref-1 ectodomain alone inhibits adipocyte differentiation indicates that generation of the inhibitory signal does not require the pref-1 cytoplasmic region. This suggests that soluble pref-1 acts as a signalling molecule through an as yet unidentified receptor. It is highly unlikely that pref-1 acts through the EGF receptor. Not only are the spacing and conservation of amino acids required for EGF-receptor interaction (38) absent in pref-1, but we have failed to detect for pref-1 the mitogenic effect normally associated with EGF receptor function (6). We hypothesize the existence of an EGF repeat containing receptor for pref-1 that could be analogous in action to the Notch-Delta receptor-ligand pair.

Although we address here the role of soluble pref-1 in adipocyte differentiation, other findings point to a broader role for pref-1 in differentiation and development. pref-1 is detected in various tissues early in embryogenesis but not in corresponding adult tissues (41). Expression of the pref-1 homolog dlk has been linked to small cell lung carcinoma and neuroendocrine tumors (29). In the larger context we hypothesize that pref-1 may maintain undifferentiated states in a number of cell types, and its downregulation may be required for differentiation. The expression of alternately spliced forms of pref-1, each with potentially distinct functions and ranges of action, may be temporally and/or spatially restricted during development. The finding that FA1, the pref-1 extracellular domain, is present in fetal circulation supports a broader in vivo role for the processing and effects of soluble pref-1 than we have described. Along these lines it is tempting to speculate that soluble pref-1 may repress adipogenesis in vivo, a process that, depending on the species, occurs late in gestation or neonatally.

ACKNOWLEDGMENTS

We thank R. Evans for the PPAR γ cDNA, S. McKnight for the C/EBP α cDNA, and S. Patel for technical assistance.

This work was supported by grant DK49620 from the National Institutes of Health to H.S.S.

REFERENCES

- Appella, E., I. T. Weber, and R. B. 1988. Structure and function of epidermal growth factor-like regions in proteins. *FEBS Lett.* 231:1-12.
- Brachmann, R., P. B. Lindquist, M. Nagashima, W. Kohr, T. Lipari, M. Napier, and R. Derynck. 1993. Transmembrane TGF- α precursors activate EGF/TGF- α receptors. *Cell* 56:691-700.
- Breyer, J. A., and S. Cohen. 1990. The epidermal growth factor precursor isolated from murine kidney membranes. Chemical characterization and biological properties. *J. Biol. Chem.* 265:16564-16570.
- Buxbaum, J. D., S. E. Gandy, P. Ciceretti, M. E. Ehrlich, A. J. Czerwik, R. P. Fracasso, T. V. Ramabhadran, A. J. Unterbeck, and P. Greengard. 1990. Processing of Alzheimer beta/A4 amyloid precursor protein: modulation by agents that regulate protein phosphorylation. *Proc. Natl. Acad. Sci. USA* 87:6003-6006.
- Carpenter, G., and S. Cohen. 1990. Epidermal growth factor. *J. Biol. Chem.* 265:7709-7712.
- Chen, L., and H. S. S. 1997. Unpublished data.
- Cheng, H. J., and J. G. Flanagan. 1994. Transmembrane kit ligand cleavage does not require a signal in the cytoplasmic domain and occurs at a site dependent on spacing from the membrane. *Mol. Biol. Cell* 5:943-953.
- Christy, R. J., V. W. Yang, J. M. Ntambi, D. E. Geiman, W. H. Landschulz, A. D. Friedman, Y. Nakabeppu, T. J. Kelly, and M. D. Lane. 1989. Differentiation-induced gene expression in 3T3-L1 preadipocytes: CCAAT/enhancer binding protein interacts with and activates the promoters of two adipocyte-specific genes. *Genes Dev.* 3:1323-1335.
- Evans, G. I., G. K. Lewis, G. Ramsey, and J. M. Bishop. 1985. Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Mol. Cell Biol.* 5:3610-3616.
- Faust, I. M., P. R. Johnson, J. S. Stern, and J. Hirsch. 1978. Diet-induced adipocyte number increase in adult rats: a new model of obesity. *Am. J. Physiol.* 235:E279-E286.
- Fehon, R. G., P. J. Koob, I. Kechay, C. L. Regan, T. Xu, M. A. T. Muskhavich, and S. Artavanis-Tsakonas. 1990. Molecular interactions between the protein products of the neurogenic loci *Notch* and *Delta*, two EGF-*homologous* genes in *Drosophila*. *Cell* 61:523-534.
- Flanagan, J. G., D. C. Chan, and P. Leder. 1991. Transmembrane form of the kit ligand growth factor is determined by alternative splicing and is missing in the *Scf* mutant. *Cell* 64:1035-1045.
- Fornace, B. M., P. Tontoz, J. Chen, R. P. Brann, B. M. Spiegelman, and R. M. Evans. 1995. 15-Deoxy- Δ^2 , 14-prostaglandin Δ^2 is a ligand for the adipocyte determination factor PPAR γ . *Cell* 83:803-812.
- Green, H., and O. Kehinde. 1976. Spontaneous heritable changes leading to increased adipocyte conversion in 3T3 cells. *Cell* 7:105-113.
- Gross, H., and O. Kehinde. 1979. Formation of normally differentiated subcutaneous fat pads by an established preadipose cell line. *J. Cell. Physiol.* 101:169-171.
- Gullberg, U., M. Lantz, L. Lindvall, I. Olsson, and A. Himmler. 1992. Involvement of an Arg-Val cleavage site in the production of a soluble form of a human tumor necrosis factor (TNF) receptor. Site-directed mutagenesis of a putative cleavage site in the p55 TNF receptor chain. *Eur. J. Cell Biol.* 58:307-312.
- Halaas, J. L., K. S. Gajiwala, M. Maffei, S. L. Cohen, B. T. Chait, D. Rabinowitz, R. L. Lallone, S. K. Burley, and J. M. Friedman. 1995. Weight-reducing effects of the plasma protein encoded by the obese gene. *Science* 269:543-546.
- Hatsuzawa, K., K. Murakami, and K. Nakayama. 1992. Molecular and enzymatic properties of furin, a Kex2-like endoprotease involved in precursor cleavage at Arg-X-Lys/Arg-Arg sites. *J. Biochem.* 111:296-301.
- Herrera, R., H. S. Ro, G. S. Robinson, K. G. Xanthopoulos, and B. M. Spiegelman. 1989. A direct role for C/EBP and the AP-1-binding site in gene expression. *Mol. Cell Biol.* 9:5331-5339.
- Hu, E., P. Tontoz, and B. M. Spiegelman. 1995. Transdifferentiation of myoblasts by the adipogenic transcription factors PPAR γ gamma and C/EBP alpha. *Proc. Natl. Acad. Sci. USA* 92:9856-9860.
- Huang, F., K. H. Neeb, J. Buck, and P. Besmer. 1992. Differential expression and processing of two cell associated forms of the kit-ligand: KL-1 and KL-2. *Mol. Biol. Cell* 3:349-362.
- Ibrahim, A., F. Bonino, S. Bardon, G. Allhand, and C. Dani. 1992. Essential role of collagens for terminal differentiation of preadipocytes. *Biochem. Biophys. Res. Commun.* 187:1314-1322.
- Jensen, C. H., T. N. Krogh, F. Holrup, P. P. Clausen, K. Skjold, L. I. Larsson, J. J. Englund, and B. Telsen. 1994. Protein structure of fetal integrin 1 (FA1). A novel circulating human epidermal-growth-factor-like protein expressed in neuroendocrine tumors and its relation to the gene products of *dlk* and *p52*. *Eur. J. Biochem.* 228:83-92.
- Kiewer, S. A., J. M. Lenhard, T. M. Wilson, I. Patel, D. C. Morris, and J. M. Lehmann. 1995. A prostaglandin Δ^2 metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation. *Cell* 83:813-819.
- Klyde, B. J., and J. Hirsch. 1979. Increased cellular proliferation in adipose tissue of adult rats fed a high-fat diet. *J. Lipid Res.* 20:705-715.
- Klyde, B. J., and J. Hirsch. 1979. Isotopic labeling of DNA in rat adipose tissue: evidence for proliferating cells associated with mature adipocytes. *J. Lipid Res.* 20:691-704.
- Kohn, T., M. T. Brewer, S. L. Baker, P. E. Schwartz, M. W. King, K. E. Hale, R. C. Thompson, R. C. Thompson, and J. L. Vannier. 1990. A second tumor necrosis factor receptor gene product can be used a naturally occurring tumor necrosis factor inhibitor. *Proc. Natl. Acad. Sci. USA* 87:8331-8335.

28. Kocyński, C. C., A. K. Allen, K. Fichtel, P. J. Kesh, and M. A. T. Muskavitch. 1988. Delta, a *Drosophila* neurogenic gene, is transcriptionally complex and encodes a protein related to blood coagulation factors and epidermal growth factor of vertebrates. *Genes Dev.* 2:1723-1735.
29. Laborda, J., E. A. Sansville, T. Hoffman, and V. Notario. 1993. Dlk, a putative mammalian homeotic gene differentially expressed in small cell lung carcinoma and neuroendocrine tumor cell line. *J. Biol. Chem.* 268:3817-3820.
30. Luettele, N. C., G. K. Michalopoulos, J. Teixido, R. Gilmore, J. Massague, and D. C. Lee. 1988. Characterization of high molecular weight transforming growth factor alpha produced by rat hepatocellular carcinoma cells. *Biochemistry* 27:6488-6494.
31. Massague, J. 1990. Transforming growth factor- α . A model for membrane-anchored growth factors. *J. Biol. Chem.* 265:21393-21396.
32. Massague, J., and A. Pandiella. 1993. Membrane-anchored growth factors. *Annu. Rev. Biochem.* 62:515-541.
33. Mroczkowski, B., M. Reich, K. Chen, G. I. Bell, and S. Cohen. 1989. Recombinant human epidermal growth factor precursor is a glycosylated membrane protein with biological activity. *Mol. Cell. Biol.* 9:2771-2778.
34. Pandiella, A., M. W. Bosenberg, E. J. Huang, P. Besmer, and J. Massague. 1992. Cleavage of membrane-anchored growth factors involves distinct protease activities regulated through common mechanisms. *J. Biol. Chem.* 267:24028-24033.
35. Parfies, G. K. Chen, K. S. Misone, and S. Cohen. 1995. The human urinary epidermal growth factor (EGF) precursor. Isolation of a biologically active 160-kilodalton heparin-binding pro-EGF with a truncated carboxyl terminus. *J. Biol. Chem.* 270:27954-27960.
36. Pellegri-Mount, M. A., M. J. Cullen, M. B. Baker, R. Hecht, D. Winters, T. Boone, and F. Collins. 1995. Effects of the obese gene product on body weight regulation in ob/ob mice. *Science* 269:540-543.
37. Rall, L. B., J. Scott, G. I. Bell, R. J. Crawford, J. D. Penschow, H. D. Niall, and J. P. Coghlan. 1985. Mouse prepro-epidermal growth factor synthesis by the kidney and other tissues. *Nature* 313:228-231.
38. Ray, P., F. J. Moy, G. T. Montellone, J. F. Liu, S. A. Narang, H. A. Scheraga, and R. Wu. 1988. Structure-function studies of murine epidermal growth factor: expression and site-directed mutagenesis of epidermal growth factor gene. *Biochemistry* 27:7289-7295.
39. Rubin, C. S., A. Hirsch, C. Fung, and O. M. Rosen. 1978. Development of hormone receptors and hormonal responsiveness *in vitro*. Insulin receptors and insulin sensitivity in the preadipocyte and adipocyte forms of 3T3-L1 cells. *J. Biol. Chem.* 253:7570-7578.
40. Smas, C. M., D. Green, and H. S. Sul. 1994. Structural characterization and alternate splicing of the gene encoding the preadipocyte EGF-like protein pref-1. *Biochemistry* 33:9257-9265.
41. Smas, C. M., and H. S. Sul. 1993. Pref-1, a protein containing EGF-like repeats, inhibits adipocyte differentiation. *Cell* 73:725-734.
42. Smas, C. M., and H. S. Sul. 1995. Control of adipocyte differentiation. *Biochem. J.* 309:697-710.
43. Smas, C. M., L. Chen, and H. S. Sul. Unpublished data.
44. Smas, C. M., S. Fong, and H. S. Sul. Unpublished data.
45. Spiegelman, B. M., and C. A. Ginty. 1983. Fibronectin modulation of cell shape and lipogenic gene expression in 3T3-adipocytes. *Cell* 35:657-666.
46. Teixido, J., and J. Massague. 1988. Structural properties of a soluble bioactive precursor for transforming growth factor- α . *J. Biol. Chem.* 263:3924-3929.
47. Tentenez, P., E. Hu, R. A. Graves, A. I. Budavari, and B. M. Spiegelman. 1994. mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev.* 8:1224-1234.
48. Wharton, K. A., K. M. Johansen, T. Xu, and S. Artavanis-Tsakonas. 1985. Nucleotide sequence from the neurogenic locus notch implies a gene product that shares homology with proteins containing EGF-like repeats. *Cell* 43:567-581.
49. Wang, S. T., L. F. Winshell, B. K. McCune, H. S. Earp, J. Teixido, J. Massague, B. Herman, and D. C. Lee. 1989. The TGF- α precursor expressed on the cell surface binds to the EGF receptor on adjacent cells, leading to signal transduction. *Cell* 56:495-506.
50. Wu, Z., Y. Xie, N. L. R. Bucher, and S. R. Farmer. 1995. Conditional ectopic expression of C/EBP β in NIH-3T3 cells induces PPAR gamma and stimulates adipogenesis. *Genes Dev.* 9:2350-2363.
51. Zhang, Y., R. Proenca, M. Maffei, M. Barone, L. Leopold, and J. M. Friedman. 1994. Positional cloning of the mouse obese gene and its human homologue. *Nature* 372:425-432.